A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2017 117

# De-duplication using automated face recognition: a mathematical model and all babies are equally cute

Luuk Spreeuwers [1]

**Abstract:** De-duplication is defined as the technique to eliminate or link duplicate copies of repeating data. We consider a specific de-duplication application where a subject applies for a new passport and we want to check if he possesses a passport already under another name. To determine this, a facial photograph of the subject is compared to all photographs of the national database of passports. We investigate if state of the art facial recognition is up to this task and find that for a large database about 2 out of 3 duplicates can be found while few or no false duplicates are reported. This means that de-duplication using automated face recognition is feasible in practice. We also present a mathematical model to predict the performance of de-duplication and find that the probability that $k$ false duplicates are returned can be described well by a Poisson distribution using a varying, subject specific false match rate. We present experimental results using a large database of actual passport photographs consisting of 224 000 images of about 100 000 subjects and find that the results are predicted well by our model.

**Keywords:** De-duplication, face recognition, large database, binomial distribution

## 1 Introduction

De-duplication is defined as the technique to eliminate or link duplicate copies of repeating data. In biometrics, there are several applications for de-duplication. One application is the cleaning of databases to make sure there is only one record per subject. A second application is to prevent that a new sample is entered in the database as a new entry, while a record of the subject already exists. In this paper, we address the 2nd category and more specifically, the application where a person applies for a new passport. The aim is to detect if this person already has a passport under another name. Currently, in the Netherlands, there exists a highly secured database of approximately 20 million subjects. The aim of this research was to investigate if it is feasible to, using modern state of the art automated facial recognition, determine if a subject has an entry in the database under another name. The main challenge in this context is the size of the database. In order to make the de-duplication feasible, if the photograph of an applicant is compared to the complete database, this should result in few to no false duplicates, caused by so-called look-a-likes, and should return true duplicates with a high probability. De-duplication becomes feasible if in 7-9 out of 10 applications, no false duplicates would be generated, while in 99 out of 100 applications the number of false duplicates would be less than 10. The latter means that an official has to manually inspect up to 10 returned images from the database

[1] Biometric Pattern Recognition Group, Chair of Services, Cyber Security and Safety (SCS), Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), University of Twente, Netherlands, l.j.spreeuwers@utwente.nl

to decide if they are actual duplicates or are caused by look-a-likes. Further in order to be effective, the probability to detect actual duplicates should at least be above 50% (every second duplicate detected). These requirements were drafted in consultation with the Dutch passport issuance institution as realistic requirements.

There is not much literature available on de-duplication in face biometrics. In [DR13], an investigative study is presented on de-duplication errors. Two types of errors are introduced: False de-duplication (FDD) which is a match with a look-a-like and False non-duplication (FND) which corresponds to a missed duplicated. They provide results on a database with 1 009 identities. In [Ya11], de-duplication based on facial feature points is reported on a database of Chinese ID cards with 60 000 entries and 100/100 duplicates detected with 8 false hits. The main subject of the paper is, however, the presentation of a face recognition method based on 105 facial feature points, and the part on de-duplication performance is very brief. Scalability is not investigated at all. There are some reports on the related subject of large-scale 1:N comparison, see e.g. [GP04, GN14], but they do not explicitly address de-duplication.

One of the aims of our research is to investigate scalability to large databases of millions of entries. The following research questions were therefore formulated:

1. Is S.O.T.A. automated face recognition good enough to reliably detect duplicates in database with a size of 20 million entries?
2. What are the settings and further requirements for effective du-duplication?
3. Can the performance of de-duplication be predicted using a model?

In order to answer these questions, we developed a model for the de-duplication performance based on the binomial and Poisson distributions and set up an experiment using a database with approximately 100 000 subjects and 230 000 images and two commercial, state of the art automated facial recognition systems.

The remainder of this paper consists of the following sections: in section 2 a mathematical model is presented that describes the probability on errors and the probability to detect duplicates in large databases. In section 3, an experiment using a large database of 100 000 subjects is presented to verify the model. Finally, conclusions are presented in section 4.

## 2 A mathematical model for detection of duplicates

### 2.1 Errors in common biometric systems

In its basic form, a biometric system compares two biometric traces, e.g. facial images, and produces a similarity score $s$ that is higher if the images are more similar. The aim of the biometric system is to determine if the two traces originate from the same the same subject. The similarity score is compared to a threshold $T$ and if $s \geq T$, the traces are

| Trace origins | result | type of match |
|---|---|---|
| Same subject | $s < T$ | False Non Match (FNM) |
| Same subject | $s \geq T$ | True Match (TM) |
| Distinct subjects | $s < T$ | True Non Match (TNM) |
| Distinct subjects | $s \geq T$ | False Match (FM) |

Tab. 1: Types of matches of a biometric comparison

classified as coming from the same subject if not, they are regarded as traces from two different subjects. For a comparison 4 cases can be distinguished as shown in Table 1.

The performance of a biometric system is represented by an ROC graph, which shows the True Match Rate (TMR) as a function of the False Match Rate (FMR) for varying threshold. The ROC shows the trade-off between the TMR and the FMR: if the FMR decreases, then the TMR decreases and if the FMR increases, then the TMR also increases. If we choose $T$ such that a certain FMR is realised, then from the ROC, we can read the TMR of the face comparison system. This is important for biometric systems that are used for verification applications, e.g. at border control where the one trace is the digital photograph stored in the passport and the other is a live recorded image. If the comparison results in a score higher than the given threshold, the probability that this is a True Match is estimated by the TMR and the probability on a False Match is estimated by the FMR, and both can be read from the ROC. The ROC is typically obtained using a large dataset of facial images.

An example of an ROC is given in Figure 1.

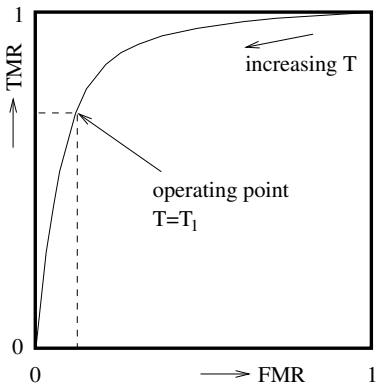


Fig. 1: ROC with operating point

A second common application of biometrics systems is the identification setting, where a single trace is compared to a list of traces of multiple subjects to check if the trace belongs to one of the subjects. We distinguish open set and closed set identification. In the former it is not known whether the owner of the trace is in the list of subjects, whereas in the latter case it is. Results are reported in the form of rank identification rates, where the rank-1 identification rate is an estimate of the probability that the subject in the list that results in

the highest score is the correct subject and rank-n that the correct subject is among the n highest scoring subjects in the list. In open set identification, also FNMR is reported and is also called False Negative Identification Rate (FNIR). Identification performance depends highly on the number of subjects in the list.

### 2.2 Performance of de-duplication

In [DR13], two types of de-duplication errors are distinguished: false de-duplication (FDD), i.e. the case that a duplicate is found while the corresponding trace in the database is actually not of the same subject as the probe trace, and false non-duplication (FND) where a trace of the same subject as the probe trace is present in the database, but not detected. These, however, apply to the case where one wants to build a database free of duplicates.

In our case, we want to detect duplicates of a facial photograph for a new passport application in a database. In order to make this feasible, we need to know the probability that a true duplicate (TD) is detected and the probability that the number of false duplicates (NFD) is below a certain threshold. For this we introduce the following measures:

| Description | measure |
|---|---|
| Probability that a true duplicate is detected | $P(\text{TD})$ |
| Probability on $k$ false duplicates | $P(\text{NFD} = k)$ |
| Probability that number of false duplicates is less than $k$ | $P(\text{NFD} < k)$ |

Tab. 2: Measures for de-duplication, TD=True Duplicate, NFD=Number of False Duplicates

In the introduction we suggested that de-duplication is feasible in practice in the passport application if $P(\text{TD}) > 0.5$, $P(\text{NFD} = 0) > 0.7$ and $P(\text{NFD} < 10) > 0.99$.

### 2.3 A mathematical model for de-duplication

We assume that we have a facial image of a subject X and a large dataset of $M$ images of which there are $N_D$ duplicates and $N$ images of other subjects. Furthermore, we assume that we have an automated face recognition (FR) system that compares two images, resulting in a score that is compared to a threshold $T$. The performance of the FR system is defined by its ROC, i.e. for a threshold $T$, we know the corresponding TMR and FMR.

If we compare the trace of X to all images in the database, then the probability that we detect a specific duplicate is given by the probability of a true match ($\alpha$) when the trace is compared to a duplicate, i.e. it is estimated by the TMR obtained from the ROC.

$$P(\text{TD}) = \alpha \approx \text{TMR} \tag{1}$$

The probability on $k$ false duplicates is modelled by a a series Bernoulli trials, where the probability on a false duplicate for a single comparison ($\beta$)is estimated by the FMR. The probability on $k$ false duplicates is then given by the binomial distribution:

$$P(\text{NFD} = k) = \binom{n}{k} \beta^k (1-\beta)^{N-k} \tag{2}$$

This is the probability that $k$ comparisons result in a score above $T$, while $N-k$ result in a score below $T$. The probability that less than $k$ false duplicates are detected is then:

$$P(\text{NFD} < k) = \sum_{i=0}^{k-1} \binom{n}{k} \beta^k (1-\beta)^{N-k} \tag{3}$$

Note that an 1:N comparison is in practice not always described properly by N 1:1 comparisons, because FR systems may use various ways of score normalisation. For our derivations we ignore this effect.

Now, it can be shown that if $N$ is very large and $N >> k$, then the binomial distribution can be approximated by the Poisson distribution [PP02]:

$$P(\text{NFD} = k) = \binom{n}{k} \beta^k (1-\beta)^{N-k} \approx \frac{1}{k!} \mu^k e^{-\mu} \tag{4}$$

Here, $\mu = N\beta$. Now this has an interesting implication if we want to predict the behaviour of de-duplication for varying database size $N$. If $N$ increases by a factor $\lambda$, then if at the same time $\beta$ (or the FMR) is decreased by a factor $\frac{1}{\lambda}$, the same probabilities result for $P(\text{NFD} = k)$ and $P(\text{NFD} < k)$!

The Poisson distribution has three different modes, depending on $\mu$:

| range of $\mu$ | behaviour as a function of $k$ |
|---|---|
| $\mu \leq 1$ | strictly decreasing |
| $1 < \mu \leq 5$ | first going up, then down |
| $5 < \mu$ | starting at nearly 0 going up then down |

Tab. 3: Behaviour of the Poisson distribution as a function of $\mu$

The three modes are also illustrated in Figure 2. Note that since k is an integer, the curves are not continuous.

Since we require $P(\text{NFD} = 0) > 0.7$, we need $\mu < 0.5$. As a matter of fact, we can calculate $P(\text{NFD} = 0)$ as a function of $\mu$ and likewise $P(\text{NFD} < k)$ as well. These relations are shown in Figure 3, where in the right figure $1 - P(\text{NFD} \leq 10)$ is plotted.

We can derive that for $P(\text{NFD} = 0) > 0.7$, we need $\mu < 0.36$, for $P(\text{NFD} = 0) > 0.9$, we need $\mu < 0.11$ and for all $\mu < 2$, $P(\text{NFD} < 10) >> 0.99$. Since $\mu = N\beta$, we can also

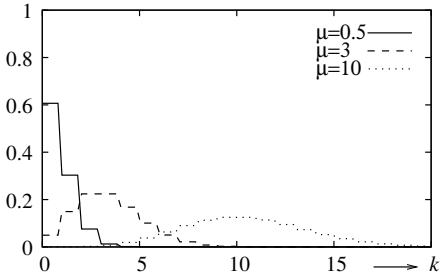


Fig. 2: Poisson distribution for various $\mu$

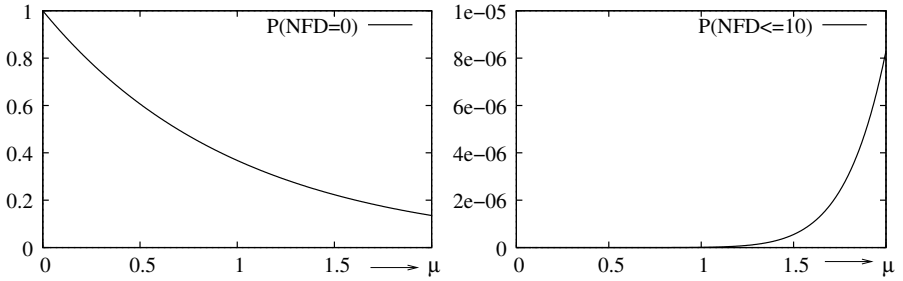


Fig. 3: $P(\text{NFD} = 0)$ and 1-$P(\text{NFD} \leq 10)$ as a function of $\mu$

calculate the required $\beta$ or FMR for a given dataset size. For various dataset sizes the required FMR values are given in Table 4.

| $N$ | $\beta$ for $P(\text{NFD} = 0) = 0.9$ | $\beta$ for $P(\text{NFD} = 0) = 0.7$ |
|---|---|---|
| 1 000 | $1.1 \cdot 10^{-4}$ | $3.6 \cdot 10^{-4}$ |
| 100 000 | $1.1 \cdot 10^{-6}$ | $3.6 \cdot 10^{-6}$ |
| 200 000 | $5.5 \cdot 10^{-7}$ | $1.8 \cdot 10^{-6}$ |
| 10 000 000 | $1.1 \cdot 10^{-8}$ | $3.6 \cdot 10^{-8}$ |
| 20 000 000 | $5.5 \cdot 10^{-9}$ | $1.8 \cdot 10^{-8}$ |

Tab. 4: Required $\beta$ or FMR for various dataset sizes

In conclusion, we can state that it is very well possible to predict the large scale behaviour of de-duplication using the Poisson distribution. There is, however, one catch: when we model the distribution $P(\text{NFD} = k)$ using the binomial distribution with constant $\beta$, we assume that for every subject, this $\beta$ (or FMR) is the same. This, however, is not the case: some subjects are easier recognised than others and some subjects look more like each other than others. The used $\beta$ is actually only the *average* $\beta$, $\bar{\beta}$ over all subjects. Thus $\beta$ will vary per subject. In order to investigate the dependency of the results on the variation of $\beta$, we assumed that $\beta$ would vary between $0.1\bar{\beta}$ and $1.9\bar{\beta}$ with a homogeneous distribution.

The probability on a certain number of false duplicated is thus calculated as:

$$P(\text{NFD} = k) = \int_{0.1\bar{\mu}}^{1.9\bar{\mu}} \frac{1}{k!}\mu^k e^{-\mu} d\mu \tag{5}$$

Where $\bar{\mu} = N\bar{\beta}$. Of course this is not the actual distribution of $\beta$, but it at least gives an indication of the effect of varying $\beta$ for the different subjects. In Figure 4 the effect of varying $\mu$ (same as varying $\beta$, since $\mu = N\beta$) is shown.

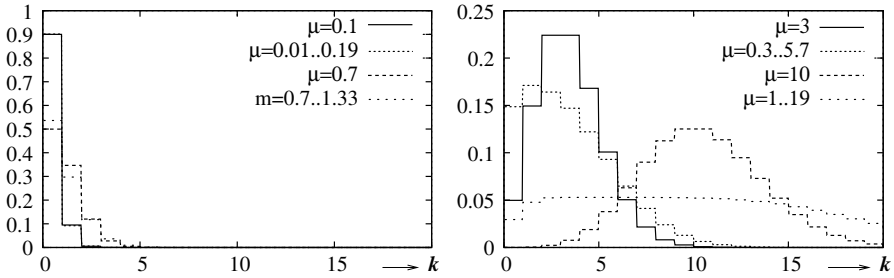


Fig. 4: Effect of non constant $\mu$ on $P(\text{NFD} = k)$, where $0.1\bar{\mu} < \mu < 1.9\bar{\mu}$

We can observe that for $\bar{\mu} = 0.1$, the effect is negligible (the curves for constant and varying $\mu$ coincide), for $\bar{\mu} = 0.7$ the peak at $k = 0$ is shifted up slightly and the tail becomes slightly longer. For larger $\bar{\mu}$, the peak of the curve $P(\text{NFD} = k)$ shifts to the left, while the whole curve becomes flatter and the right tail is longer.

Since we are interested in values of $\mu$ in the order of 0.1, we may expect that the subject specific variation in $\beta$ has only small impact on the number of expected false duplicates.

## 3 An experiment on passport data

We set up an experiment with a database of passport photographs that was made available by the Ministry of Interior and Kingdom Relations of the Netherlands. Since strict privacy regulations apply to this database, the data could only be accessed in a highly secured environment and were only available for generating comparison scores and to a limited extend for visual inspection. In total the database consisted of 224 000 images of approximately 100 000 subjects. Of most subjects only two images were available, but of some more.

Using 2 commercial face recognition (FR) systems, all images of all subjects were compared to all other images, which would result in $50 \cdot 10^9$ scores. Due to time and space limitations, fewer scores were calculated. For the first system, 217 049 and for the second system 101 000 images were compared to all 224 000 images.

First the ROC for both FR systems were determined. They are not provided here, because their shape may reveal their origin. From Table 4, we can read the required FMRs ($\beta$)

that for databases of 200 000 and 20 000 000 images. For these settings the two facial recognition systems have a TMR as reported in Table 5.

| Dataset size | $P(\text{NFD}=0)$ | FMR | TMR system 1 | TMR system 2 |
|---|---|---|---|---|
| 200 000 | 0.9 | $5.5\cdot10^{-7}$ | 0.76 | 0.82 |
| 200 000 | 0.7 | $1.8\cdot10^{-6}$ | 0.79 | 0.84 |
| 20 000 000 | 0.9 | $5.5\cdot10^{-9}$ | 0.23 | 0.22 |
| 20 000 000 | 0.7 | $1.8\cdot10^{-8}$ | 0.56 | 0.51 |

Tab. 5: FMR and TMR for two FR systems

From Table 5, we can see that for a dataset size of 200 000 the systems perform quite reasonably and allow for around 80% of the duplicates to be detected (4 out of 5). However, for a dataset of 20 000 000 the probability on detection a true duplicate drops to barely above 50% if $P(\text{NFD}=0)=0.7$. Note that with a FMR of $5.5\cdot10^{-9}$ we are at the limit of statistical certainty, because we have only about $20-40\cdot10^{9}$ false positive scores available. Also some subjects had a very high number of false duplicates, upto a few hundreds. Therefore, we visually inspected the images of the concerning subjects. To our surprise, they appeared to be all of babies and toddlers and young children, see Figure 5. As one of the results of this research we can therefore state that all babies look equally cute for the used FR systems. Indeed, poorer performance of FR for children has been reported before, see e.g. [GN14].

Fig. 5: All babies are equally cute (images obtained from the www)

We repeated the experiment with only subjects of ages above 14 years old, the results of which are represented in Table 6.

| Dataset size | $P(\text{NFD}=0)$ | FMR | TMR system 1 | TMR system 2 |
|---|---|---|---|---|
| 200 000 | 0.9 | $5.5\cdot10^{-7}$ | 0.89 | 0.92 |
| 200 000 | 0.7 | $1.1\cdot10^{-6}$ | 0.92 | 0.94 |
| 20 000 000 | 0.9 | $5.5\cdot10^{-9}$ | 0.28 | 0.27 |
| 20 000 000 | 0.7 | $1.1\cdot10^{-8}$ | 0.65 | 0.65 |

Tab. 6: FMR and TMR for two FR systems for subjects with age 14+

We now see that for a database size of 20 000 000, 7 out of 10 subjects return no false duplicates and almost 2 out of 3 true duplicates are found according to our mathematical model, which, according to our set criteria is acceptable.

To investigate if the mathematical model is valid, we compared the predicted behaviour at various settings with the actual behaviour. From the complete set of 224 000 images, we drew 3 sets of 100 000, 10 000, and 1 000 images respectively and determined the probability on $k$ false duplicates for a FMR such that $\mu = N\cdot\text{FMR} = 0.1$ (Figure 6 on

the left), and $\mu = N \cdot \text{FMR} = 1$ (Figure 6 right). We also predicted the behaviour with the models described in equations 4 and 5. These are shown as the solid curves in Figure 6.

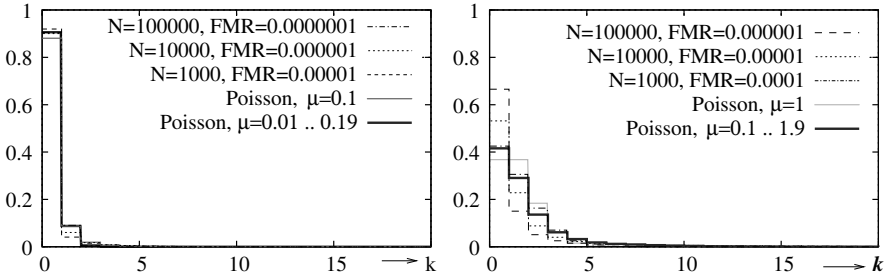


Fig. 6: Comparison of predictions by the mathematical model with actual measurements; for small $\mu$, the model (drawn lines) match the measured results (various dashed/dotted lines) very well, while for larger $\mu$ the deviations are bigger

From the curves in Figure 6, we can observe that for small $\mu$ (left), the model predicts the behaviour very well and the behaviour for varying database sizes with fixed product $N \cdot \beta$ is replicated well. This means we can predict the behaviour for larger databases reliably. For larger $\mu$, the accuracy of the prediction is less, but still the basic behaviour is characterised quite well (figure on the right). We can also observe that the model of Equation 5 for varying $\mu$ better predicts the behaviour than the Poisson distribution (Equation 4).

## 4 Conclusion

In this article we studied a specific de-duplication application where a subject applies for a new passport and we want to check if he possesses a passport already under another name. To determine this, a facial photograph of the subject is compared to all photographs of the national database of passports, in the Netherlands with a size of about 20 000 000. We investigate if state of the art facial recognition is up to this task and find that for a database of this size, duplicates can be detected with a probability of 65% (about 2 out of 3 duplicates is detected), while in 70% of all cases no false duplicates are reported and in more that 99% of all applications fewer than 10 false duplicates. This means that de-duplication using automated face recognition is feasible in practice.

We developed a mathematical model to predict the performance of de-duplication and find that the probability that $k$ false duplicates are returned can be described well by a Poisson distribution using a varying, subject specific false match rate. An interesting and very useful property of the Poisson model is that if the database size increases $N$ with a factor $\lambda$, the same behaviour is obtained provided the threshold for the FR system is chosen such that the FMR decreases with a factor $\frac{1}{\lambda}$, i.e. the product $N \cdot \text{FMR}$ remains constant.

Finally, we found that the used FR systems cannot distinguish small infants very well: for them all baby faces are equally cute.

## References

[DR13] DeCann, B.; Ross, A.: De-duplication errors in a biometric system: An investigative study. In: 2013 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 43–48, Nov 2013.

[GN14] Grother, Patrick J.; Ngan, Mei L.: , Face Recognition Vendor Test (FRVT) Performance of Face Identification Algorithms NIST IR 8009, 2014.

[GP04] Grother, P.; Phillips, P. J.: Models of large population recognition performance. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. volume 2, pp. II–68–II–75 Vol.2, June 2004.

[PP02] Papoulis, A.; Pillai, S.U.: Probability, random variables, and stochastic processes. McGraw-Hill electrical and electronic engineering series. McGraw-Hill, 2002.

[Ya11] Yang, Xiaoli; Su, Guangda; Chen, Jiansheng; Su, Nan; Ren, Xiaolong: Large Scale Identity Deduplication Using Face Recognition Based on Facial Feature Points. In (Sun, Zhenan; Lai, Jianhuang; Chen, Xilin; Tan, Tieniu, eds): Biometric Recognition: 6th Chinese Conference, CCBR 2011, Beijing, China, December 3-4, 2011. Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 25–32, 2011.