D8.8 Document anonymization module 2

Document Due Date: 31/08/2020
Document Submission Date: 24/08/2020

Work Package 8: Document Verification

Document Dissemination Level:
Public

D4FLY - **D**etecting **D**ocument frau**D** and i**D**entity on the fly
This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 833704

Abstract

One of the principal aims of the D4FLY research project is to develop technologies that will automate the analysis of travel, identity and breeder documents in order to detect fraud. These documents contain personal data, such as name, date of birth and national number. To ensure the protection of privacy as well as compliance with the GDPR and national legislation, anonymization techniques are developed for the planned research activities in D4FLY. These anonymization techniques can also be used to ensure privacy protection during D4FLY demonstration or dissemination activities. This document describes the tool that will be used for automated anonymization of travel and breeder documents. The tool consists of a graphical user interface, document recognition, keyword detection, face detection, number detection, barcode detection and masking of personal data. The results show that only 10 annotated images are needed to reach a keyword detection accuracy of 96% and an anonymization sensitivity of 93% of the related personal data. Based on the tolerance for errors, which is typically “none”, there should always be a manual inspection of these results.

Project Information

Project Name	Detecting Document frauD and iDentity on the fly
Project Acronym	D4FLY
Project Coordinator	Veridos GmbH
Project Funded by	European Commission
Under the Programme	Horizon 2020 Secure Societies
Call	H2020-SU-SEC-2018
Topic	SU-BES02-2018-2019-2020 Technologies to enhance border and external security
Funding Instrument	Research and Innovation Action
Grant Agreement No.	833704

Document Information

Document reference	**D8.8**
Document Title	**Document anonymization modules 2**
Work Package reference	WP8 Document Verification
Delivery due date	31/08/2020 [M6]
Actual submission date	24/08/2020
Dissemination Level	Public
Author(s)	**Henri Bouma, Arthur van Rooijen, Raimon Pruim, Johan-Martijn ten Hove, Jelle van Mil (TNO).**
Reviewers	Jeroen van Rest (TNO) Ben Kromhout (IND) Alfonsas Juršėnas (BPTI) Hans de Moel (RNM) Henri Bouma (TNO; WP8 leader) Zachary Goldberg (TRI; Ethical and Legal Aspects) Dimitris Kyriazanos (NCSRD; Security Advisory Board) Armin Reuter (VD; Project Coordinator)

Document Version History

Version	Date created	Beneficiary	Comments
0.1	16/07/2020	TNO	Initial draft by the authors.
0.2	20/07/2020	TNO	New version based on reviews from IND and TNO.
0.3	24/07/2020	TNO	New version based on reviews from RNM and BPTI and TRI/WP3.
0.4	20/08/2020	TNO	New version based on review NCSRD/SAB.
1.0	21/08/2020	TNO	Final version based on review from VD/Coordinator.

List of Acronyms and Abbreviations

ACRONYM	EXPLANATION
BPTI	Baltic Institute of Advanced Technology
CNN	Convolutional Neural Network
D4FLY	Detecting Document frauD and iDentity on the fly (H2020 project)
EC	European Commission
EU	European Union
GPU	Graphical Processing Unit
GUI	Graphical User Interface
IND	Immigration and Naturalization Service (The Netherlands)
NCSRD	National Centre for Scientific Research 'Demokritos'
OCR	Optical Character Recognition
PDF417	Portable Data File 417 (2D barcode)
QR	Quick Response (2D barcode)
RNM	Royal Netherlands Marechaussee
SAB	Security Advisory Board
TNO	Netherlands Organization for Applied Scientific Research
TRL	Technology Readiness Level
TRI	Trilateral Research Ltd
VD	Veridos GmbH
VGG	Visual Geometry Group
WP	Work package

Table of Contents

1 INTRODUCTION

1.1 Background

The D4FLY project will augment the current capabilities and capacities of border guards and immigration services in countering emerging threats in document and identity verification (e.g., forged documents, impostor fraud, morphed faces) at manual and highly automated border control points (both in first and in second line) and in the issuance process of genuine documents. In the project, technologies are developed that automate the analysis of travel, identity and breeder documents in order to detect fraud. Development of modules for document processing in D4FLY requires examples of these documents. Therefore, it is necessary to share the documents with the developers and use them in field tests.

These documents contain personal data of natural persons, such as their name, date of birth and national identification number. The personal data must be well protected and a data breach must be avoided at all times[1]. One of the ways to protect the personal data is to minimize the sharing of personal data (see GDPR article 5-c)[2]. Anonymization removes the personal data (e.g., by replacing the personal data by a black bar) and can therefore be used to minimize the sharing of personal data (see GDPR article 25-1)[3]. This minimization can be done in two ways.

The first way is that anonymization is applied before the document processing (Figure 1-1). Personal data is removed in an early stage and only anonymized documents are shared with developers and used during field tests. The effect is that border guards or immigration services can record data and anonymize the data before sharing it with researchers that develop modules for document processing. The advantage of this way is that even researchers cannot access the personal data, which is optimal for privacy protection. For some document processing modules the personal data may be irrelevant and it may be sufficient to see other details on the document.

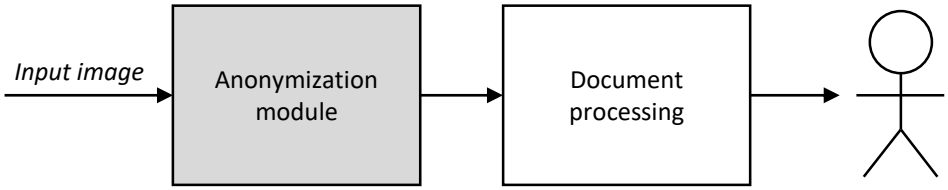


FIGURE 1-1: USE OF ANONYMIZATION BEFORE DOCUMENT PROCESSING.

[1] A data leak could lead to grave personal consequences for data subjects, and for reputation damage and serious legal repercussions for the border guard.

[2] Art 5(c), GDPR: "Personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimization')".

[3] Art 25(1): "the controller shall (…) implement appropriate technical and organizational measures, such as pseudonymization, which are designed to implement data-protection principles, such as data minimization, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects."

The second way is that anonymization is applied after document processing (Figure 1-2). In some cases the processing requires raw input images that have not been modified by an anonymization module. For example, a document authentication module that verifies whether the personal identification numbers are manipulated must be able to inspect those regions. In this case it is necessary to share the document including the personal data to the researchers, and the researchers should take appropriate actions to protect the data. Although it is necessary to process the raw input images, it may not be necessary to show the complete image on the screen. For example, during field tests, demonstration or dissemination activities, it may be sufficient to show anonymized images and thereby minimize the sharing of personal data to a restricted audience (e.g., evaluators or other end users).

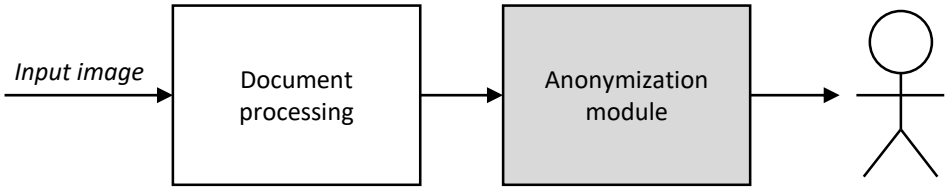


FIGURE 1-2: USE OF ANONYMIZATION AFTER DOCUMENT PROCESSING.

1.2 Recent progress

Task 8.1 in the D4FLY project focuses on anonymization and results in two deliverables: An initial deliverable D8.1 in month M6 (February 2020) and an update deliverable D8.8 in month M12 (August 2020). The differences between D8.1 and D8.8 are the following:

- **Retraining**: The current version can be retrained by the end users, while the previous version could not.
- **Rotation**: The current version applies automatically rotation before keyword detection and masking. This helps to align the boxes in a more consistent way (and reduce the over-segmentation).
- **Barcode**: The previous version of barcode detection was tested on a very small dataset (less than 10 barcodes). The new version is tested on a larger collection (41 barcodes).
- **OCR**: OCR text detection is added to improve the localization of the masks.
- **Evaluation**: the system is evaluated on two extra document types.

1.3 Aim of this document

This document describes the tool that will be used during the D4FLY project for anonymization of travel documents and breeder documents. The aim is to deliver a tool at technology readiness level (TRL) 6, so that it can be used during field tests and demonstrations in a relevant environment. Potentially, it can also be used after the project by other end users and researchers.

2 USER WORKFLOW

2.1 Introduction to the user workflow

The anonymization tool is intended to be used for travel or breeder documents that contain personal data. To avoid undesired transfer of personal data from one organization to another organization, the anonymization tool is provided with an empty database and with untrained models. Typically, the anonymization tool needs several manually annotated examples of the same document type before it can automatically anonymize that type of document.

On one hand new annotations and training are required to use the tool. On the other hand, the amount of annotations is minimized and only a few examples are needed (see Chapter 5 for results).

2.2 User procedure

The principal users will be personnel of border guard organizations and immigration services. The system has the capability to detect faces, barcodes, numbers and keywords in the scanned image of the document. This enables the system to localize personal data in the scanned image and replace the personal data by a black bar. The tool needs several manually annotated examples before it can anonymize automatically.

The interactive user procedure consists of the following steps:

1. Start the application
2. Load data (one or multiple scans)
3. Verify country and doc-type, and optionally correct
4. Verify document rotation, and optionally correct
5. Verify the anonymization, and optionally correct
6. Export the anonymized image
7. Go to next scan
8. Retrain
9. Continue (load more data) or stop.

More details can be found in the user manual (Annex A).

3 ARCHITECTURE AND INTERFACES

This chapter describes the top-level architecture and short interface description.

3.1 Top-level architecture

The top-level architecture of the anonymization module is shown in Figure 3-1. It consists of two components: an anonymization module and an anonymization graphical user interface (GUI). The remainder of this chapter will focus on the interfaces of the anonymization module. More details about the anonymization module can be found in Chapter 4 and more details about the anonymization GUI can be found in Chapter 2.

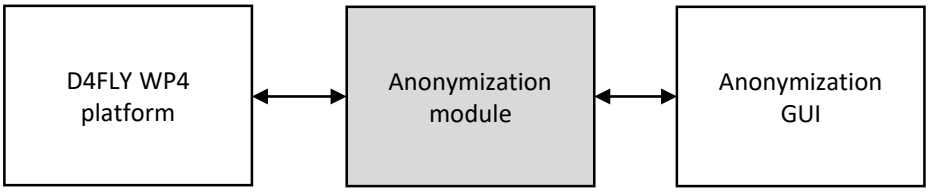


FIGURE 3-1: TOP-LEVEL ARCHITECTURE.

3.2 Short interface description

There are several interface messages between the anonymization GUI and the anonymization module. The normal workflow uses these in four steps (see Figure 3-2). There are three different interface messages (the message format of step 2 and 4 is identical).

1. In the first step, an image is submitted from the GUI to the module. This image may contain personal data.
2. In the second step, an automatic anonymization is performed and submitted from the module to the GUI.
3. In the third step, the automatic anonymization is modified and after manual verification submitted from the GUI to the module.
4. In the fourth step, the module again submits the anonymization message, which contains an update of the anonymized image. This anonymized image can be exported by the user.

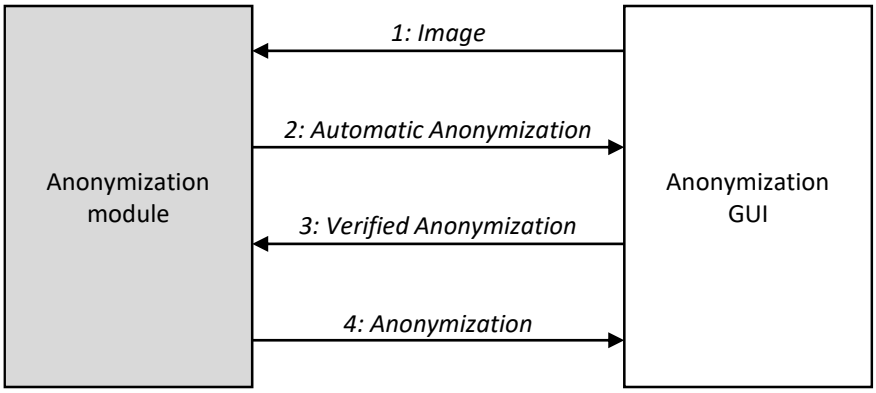


FIGURE 3-2: INTERFACES.

Automatic batch processing can be implemented in the D4FLY platform by only using step 1 and 2, without human verification.

4 METHODS

The documents may contain different types of personal data. The four most important categories are the following:

- **Photograph**: The document may contain a photograph, which reveals the facial information from a person.
- **Barcode**: The document may contain a barcode that contains personal data that is easy to read for machines.
- **Number without keyword**: The document may contain a (large) document number without any neighboring keyword.
- **Pair of keyword and value**: The document may contain pairs of keywords and values. For example:
    - Keyword = “Last name”, Value = “Smith”
    - Keyword = “Date of Birth”, Value = “31-12-2001”

Each of the categories is recognized in a different way by the anonymization module.

The Chapters 4 and 5 are a derivative work based on a paper published by the Society of Photo-Optical Instrumentation Engineers (SPIE) [Bouma, 2020].

4.1 Internal architecture

The internal architectural overview is shown in Figure 4-1.

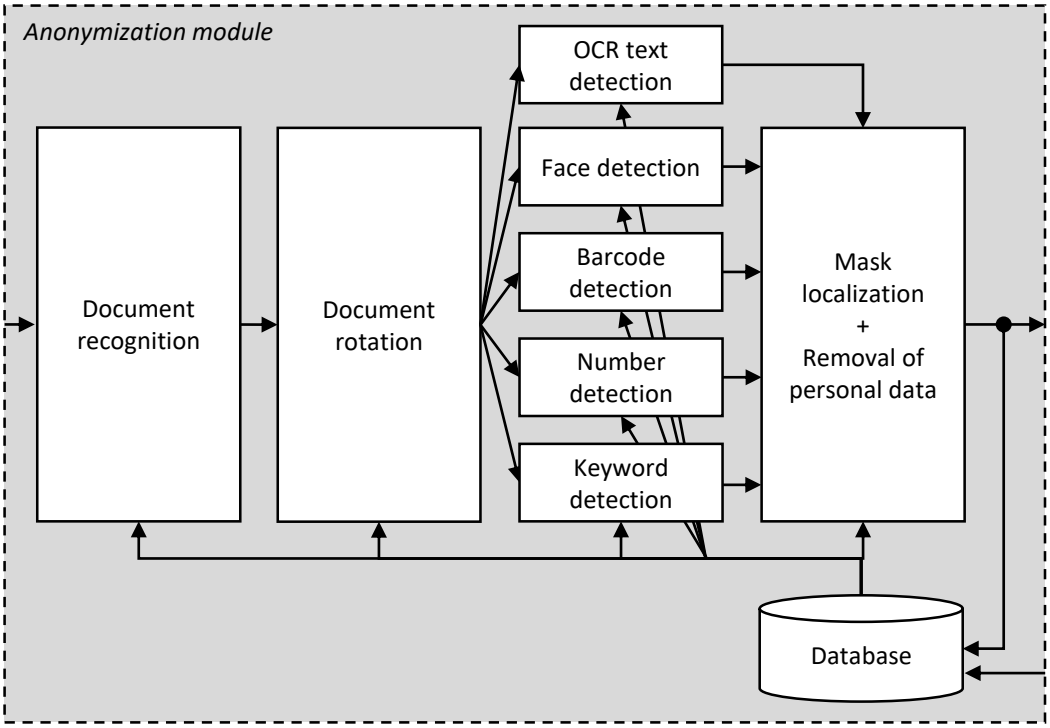


FIGURE 4-1: INTERNAL ARCHITECTURE.

4.2 Document recognition

The module for “document recognition” automatically recognizes the country name and the document type (e.g., “identity card” or “birth certificate”). The module is implemented with

VGG16 [Simonyan, 2015][VGG16], which has been chosen because it is an easy-to-use open-source deep-learning algorithm for image classification.

4.3 Document rotation

The module for "document rotation" automatically detects the document orientation and uses this calculated angle to put the document upright. This module is implemented using the Hough Line Transform and Canny Edge detector from OpenCV [OPENCV]. The calculated document angle is the median angle of all the found horizontal lines (max +/- 15 degrees) in the document.

4.4 Face detection

The module for "face detection" automatically detects and localizes a photograph. The module is implemented with a specific DLIB face detector [DLIB], which was chosen because it is a commonly-used state-of-the-art open-source algorithm for face analysis.

4.5 Barcode detection

The module for "barcode detection" automatically detects and localizes barcodes of different types (1D-barcodes or 2D-barcodes such as QR-code or PDF417). The module is implemented with a Faster R-CNN [Ren, 2017] , which is a state-of-the-art deep-learning algorithm for object detection and classification [Boer, 2017].

4.6 Number detection

The module for "number detection" can localize numbers even when there are no keywords in the neighborhood. The number detection is implemented with TESSERACT [TESSERACT_OCR], which was chosen because it one of the most accurate open-source optical character recognition (OCR) engines. The module generates multiple strings (words) on a document. If the string contains multiple digits, then the string is recognized as a number.

4.7 Keyword detection

Keywords (such as 'First name' or 'Last name' or 'Date of birth') can be detected with a specific keyword detector – which detects the keyword as a whole.

Keyword detection is implemented with a Faster R-CNN [Ren, 2017], which is a state-of-the-art deep-learning algorithm for object detection and classification [Boer, 2017]. The pre-processing consists of data augmentation by copying the relevant keyword to other locations in the same document image and therefore generating multiple artificial document images. The augmentation uses color transformation and blending.

4.8 OCR text detection

Text detection is performed using the CRAFT text detector [Baek, 2019] as implemented in the open Keras-OCR engine [KERAS-OCR]. This implementation localizes words in the images. This detector was chosen instead of TESSERACT since the CRAFT detector is considered more suitable for detecting arbitrarily-oriented and curved text. The OCR text detection is used to improve the localization of masks.

4.9 Mask localization and removal of personal data

The personal data in an image must be masked to generate anonymized images. Masks for the face, barcode and number are trivial to implement, because the location of the mask is identical to the positioning information generated by the (face, barcode or number) detection. However, the pairs of keyword and value are less trivial, because the location of the value (that contains personal data) should be derived from the detected keywords. The method is implemented in two steps.

The first step calculates the distance between every mask-keyword combination in every document image in the set of annotated and verified document images of the same country and document type. Because the document images are of varying resolutions, the distances used in this process are relative to the page height and width of the document. Some languages read from left to right and other languages read from right-to-left. In this distance calculation, the top-left coordinate (x1, y1) of the keyword is used as anchor point if the keyword is on the right side of the mask (reading right-to-left), and the bottom-right coordinate (x2, y2) otherwise. When all relative distances are calculated, the 5^{th} percentile (top-left coordinate of the mask) and 95^{th} percentile (bottom-right coordinate of the mask) are used as a 'template offset' for that mask-keyword combination.

The second step calculates the location of the masks for a new document, using the detected keywords, the template offsets and OCR text-detections:

1. For every keyword that should be in the document, it is checked if the keyword is detected. If the keyword is detected at the expected place, the direct keyword-mask offset is used. If a keyword is not detected, or if the detected keyword is not near the expected place in the document, the keyword-mask offset from the nearest keyword is used. For every country and document type, we keep a list of expected locations of keywords and masks based on the annotated and verified documents.
2. The bounding box of a mask is calculated by first retransforming the relative offset to an absolute offset and then adding the offsets to the anchor point of the detected keyword.
3. The resulting bounding box of the mask is optimized using OCR text-detections to fit the mask more closely around the text. Only the OCR text detections are valid that (1) do overlap with the mask and (2) do not overlap with the keyword and (3) with a larger width than height. The outer coordinates of these valid OCR detections are used to optimize the location of the anonymization mask. The resulting anonymization mask may become larger (e.g., if the OCR detections are partially overlapping) or smaller (e.g., if the OCR detections are inside the mask).

Finally, the pixels in the image that are located inside the masks are replaced by black pixels, thereby removing the personal data from the image.

5 EXPERIMENTS AND RESULTS

This chapter gives an overview of the experimental setup and results. The experiments focus on technical accuracy of the components for document recognition, face detection, barcode detection, keyword recognition, number detection and masking.

The Chapters 4 and 5 are a derivative work based on a paper published by the Society of Photo-Optical Instrumentation Engineers SPIE [Bouma, 2020].

5.1 Experimental setup

The anonymization tool is tested on a limited set of data. Three different document types are used and each document type has 20 scans (of 20 different documents), so the total dataset contains 60 images. In each image, faces, barcodes, numbers, keywords and personal data are marked with boxes. The used documents contain different alphabets (including Latin and Arabic) and different levels of standardization: highly standardized (identity card) and less standardized (birth/civil register).

5.2 Document recognition

Document recognition was trained on a large database with many document classes. The performance of document recognition is validated on the dataset with 60 images. The results are shown as a confusion matrix in Table 5-1, where ground-truth is indicated as 'gt' and system predictions are indicated as 'pred'. The average document recognition accuracy is 92%.

TABLE 5-1 CONFUSION MATRIX FOR DOCUMENT RECOGNITION.

	Country A (pred)	Country B (pred)	Country C (pred)	Country D (pred)	Country E (pred)
Country A (gt)	**20**	0	0	0	0
Country B (gt)	0	**20**	0	0	0
Country C (gt)	0	**8**	**12**	0	0
Country D (gt)	0	0	**0**	**20**	0
Country E (gt)	0	0	**0**	**0**	**20**

5.3 Face detection

The dataset with 60 images contains 40 faces. The performance of the pretrained face detector is validated on this dataset. The results are shown in Table 5-2, where FPR indicates the false-positive rate. The table shows that the system detects all faces without any false positives.

TABLE 5-2 RESULTS FOR FACE DETECTION.

Number of faces	Accuracy	FPR
40	100%	0%

5.4 Barcode detection

The barcode detector is trained on a part of the dataset and tested on another part of the dataset. This dataset contains 307 different files with 449 unique barcodes from 11 country and document type combinations. The barcode types are 1D, QR and PDF417. This dataset is challenging because it also contains barcodes printed on the back side of the documents. This is possible since the documents have ink on both sides. Also these barcodes can easily be missed by a human observer as they are faint. It is still important to detect and anonymize them, since they do contain sensitive information and a dedicated adversary could still extract data from them.

The detector is trained on 276 files containing 408 barcodes and evaluated on the remaining 31 documents with 41 barcodes. A barcode is considered to be detected (True Positive) if the intersection-over-union (IoU) between the detection and annotation is higher than 0.7.

The results are shown in Table 5-3, where TP is true positive, FP is false positive and FN is false negative. There are 41 barcodes, of which 40 are correctly detected (TP) and only 1 barcode is missed (FN). Furthermore, there is one false detection (FP) at a location where there is actually not a barcode present. The results show that almost 98% of the barcodes is correctly detected. We want to stress that the FP and FN are a consequence of wanting to have a network which is able to detect faint barcodes printed on the backside of the documents.

TABLE 5-3 RESULTS FOR BARCODE DETECTION.

Number of barcodes (TP+FN)	TP	FP	FN
41	40	1	1

5.5 Keyword detection

The keyword detector is trained on a part of the dataset and tested on another part of the dataset. The detector is evaluated on documents from three countries (A, B, C). A keyword is considered to be detected (True Positive) if the intersection-over-union (IoU) between the detection and annotation is higher than 0.5. The Average Precision (AP) of the detections is defined as accuracy. This detector is evaluated with leave-N-document-out cross-validation. The results are shown in Table 5-4. The results show that only 5 annotated images are needed to reach a keyword detection accuracy of 92% and 10 images are needed to reach a keyword detection accuracy of 96%.

TABLE 5-4 RESULTS FOR KEYWORD DETECTION.

Number of training images per document type	Keyword detection accuracy (AP)
1	75%
3	84%
5	92%
10	96%

5.6 Number detection

The number detector is applied to detect the document numbers in the identity cards that contain a Non-Latin document number without a related keyword. The results are shown in Table 5-5. The table shows that approximately 45% of these numbers are detected (TP) and 55% are missed (FN). The single false positive (FP) was actually not a number but part of an icon. Note that detected numbers related to a keyword (e.g., birth date) are ignored in the statistics (i.e., are considered neither as a FP or TP). The 45% is not very high, but may be acceptable for end-users during interactive annotation, because we can infer the number mask from other keyword locations.

TABLE 5-5 RESULTS FOR NON-LATIN NUMBER DETECTION.

Number of Numbers (TP+FN)	TP	FP	FN
9	4	1	5

5.7 Masking textual data

For the evaluation of the masking it is not sufficient to solely define whether the data field has been 'detected' or not. Rather, it needs to be assessed to which extent the personal data areas have been covered, and the extent of falsely anonymized areas. Therefore we define two measures for accuracy assessment of the masking:

1) True Positive Rate (i.e. sensitivity or recall): TPR = TP / (TP + FN)

2) False Positive Rate: FPR = FP / (TP + FP),

where the correct anonymization is indicated as true positive area (TP), the incorrectly missed area is indicated as false negative (FN), and the incorrectly anonymized area is indicated as false positive (FP). An intuitive illustration of TPR and FPR is shown in Figure 5-1.

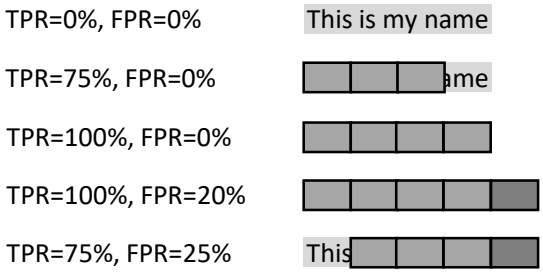


FIGURE 5-1: ILLUSTRATION OF TPR AND FPR.

The results are shown in Table 5-6. These results are obtained using the keyword detector over two train/test splits of the data for each document type (i.e., per document type train on 10 annotated images, test on the 10 remaining images, and vice versa). The table shows that the average anonymization sensitivity (TPR) is 96% when the masks are related to ground-truth keyword boxes and the average anonymization sensitivity is 93% when the masks are related to automatic keyword boxes, which means that almost all masks are placed correctly. The average FPR is 35%, which indicates some over-segmentation. For anonymization, it is better to have over-segmentation than to reveal the personal data, and therefore the trade-

off between FPR and TPR is chosen in such that the FPR is further from 0 (over segmentation) than the TPR is from 100 (missing a region with personal data).

TABLE 5-6 RESULTS FOR MASKING.

	Masking with ground-truth keyword boxes		Masking with automatic keyword detection	
	TPR	FPR	TPR	FPR
Country A	99%	34%	98%	31%
Country B	94%	31%	82%	40%
Country C	92%	44%	91%	43%
Country D	98%	33%	98%	33%
Country F	98%	34%	97%	34%
TOTAL	**96%**	**35%**	**93%**	**36%**

Figure 5-2 shows a representative example with an FPR of 35% to give an impression of the amount of over-segmentation, where the yellow boxes are the manual masks and the black boxes are the automatic masks (photo and number are white and keywords are blue).

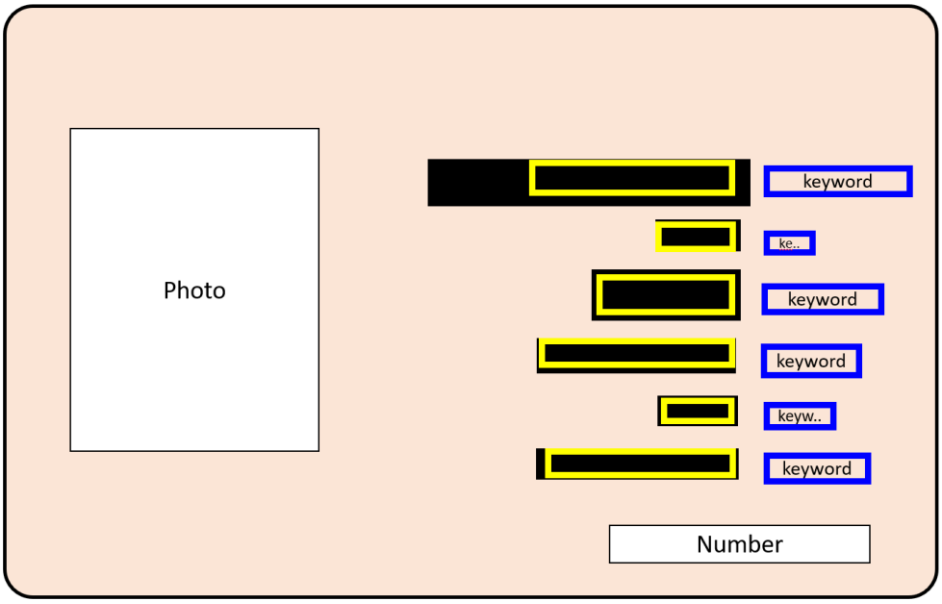


FIGURE 5-2: EXAMPLE RESULT OF MASKING TEXTUAL DATA WITH FPR=35%.

6 CONCLUSIONS

Document anonymization is important to minimize the sharing of personal data and anonymization can be used during demonstration or dissemination activities. This document describes a tool that will be used for anonymization of travel and breeder documents in the D4FLY project. The tool consists of a graphical user interface, document recognition, keyword detection, face detection, number detection, barcode detection and masking of personal data. The results show that only 10 annotated images are needed to reach a keyword detection accuracy of 96% and an anonymization sensitivity of 93% of the related personal data. Face and barcode detection also reached a high accuracy of 100% and 98% respectively. Only the OCR-based number detection has lower performance but can be compensated by approximate localization of masks based on keywords. Based on the tolerance for errors, which is typically “none”, there should always be a manual inspection of these results.

REFERENCES

[Baek, 2019]	Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. "Character Region Awareness for Text Detection," IEEE CVPR, (2019)
[Boer, 2017]	Boer M. de, Bouma, H., Kruithof, M., et al., "Automatic analysis of online image data for law enforcement agencies by concept detection and instance search," Proc. SPIE 10441, (2017).
[Bouma, 2020]	Bouma, H., Pruim, R., van Rooijen, A., ten Hove, J., van Mil, J., Kromhout, B., "Document anonymization for border guards and immigration services," Proc. SPIE 11542, (2020).
[DLIB]	http://dlib.net
[D4FLY-D8.1]	D4FLY, "D8.1: Anonymization module 1", 2020.
[D4FLY-D8.4]	D4FLY, "D8.4: Tactical anomaly detection 1", 2020.
[Hermans, 2017]	Hermans, A., Beyer, L., & Leibe, B., "In defense of the triplet loss for person re-identification," arXiv:1703.07737, (2017).
[KERAS_OCR]	https://github.com/faustomorales/keras-ocr
[OPENCV]	https://opencv.org
[PYZBAR]	https://github.com/NaturalHistoryMuseum/pyzbar
[Ren, 2017]	Ren, S., He, K., Girshick, R., Sun, J., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", IEEE Trans. Pattern Analysis and Machine Intelligence 39, 1137-1149 (2017).
[Simonyan, 2015]	Simonyan, K., & Zisserman, A., "Very deep convolutional networks for large-scale image recognition," ICLR, (2015).
[TESSERACT_OCR]	https://github.com/tesseract-ocr/
[VGG16]	https://keras.io/applications/

ANNEX A: USER MANUAL

This annex contains a step-by-step user manual. A short introduction to the user workflow can be found in Chapter 2.

1. Start the application

Use the following steps to start the application:

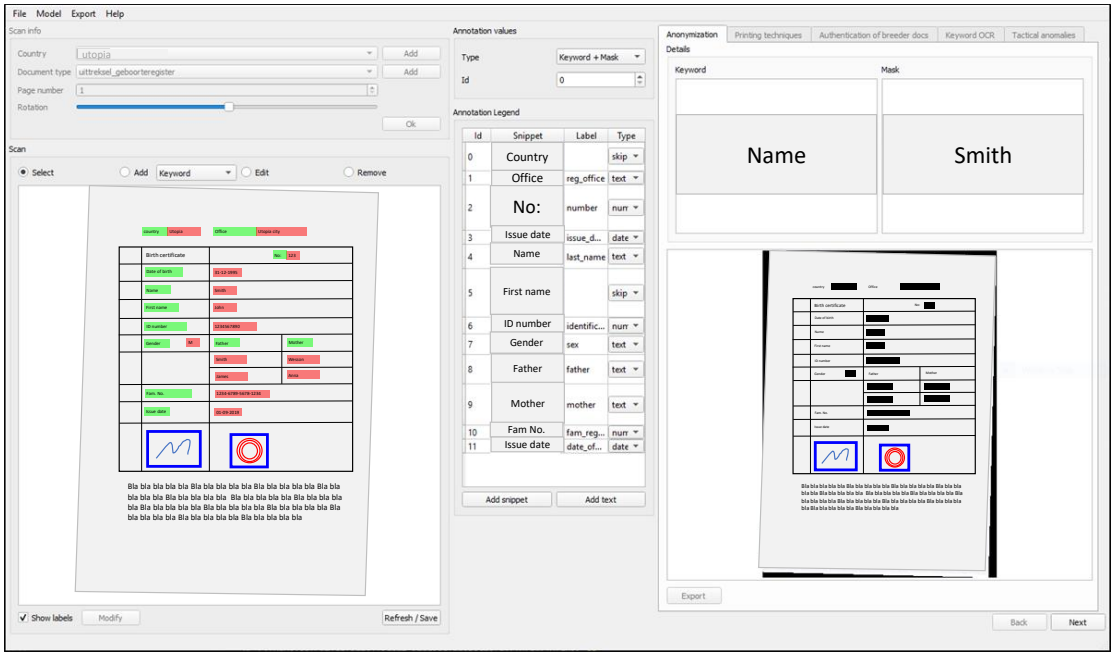
- Select on the desktop → "D4FLY_BreederAnalysis".

2. Load data (one or multiple scans)

Use the following steps to load data:

- Select on the top menu-bar → "File" → "Open files".
- Browse to the folder with input images.
- Select one or multiple files (with "shift" or "control" and the left mouse button)
- Select the "Open" button.

The first image shown on the screen. A screenshot of the graphical user interface (GUI) is shown in the figure below. The screenshot is modified to remove all information from the original scan.



3. Verify country and doc-type and rotation, and optionally correct

The system automatically recognizes "Country" and "Document type" and "Rotation".

- **Country**: The first output of document recognition is shown as "Country" and this can be modified by the user with a drop-down menu.
- **Document type**: The second output of document recognition is shown as "Document type" (e.g., Identity card, Birth certificate) and this can be modified by the user with a drop-down menu.

- **Rotation**: For anonymization and data extraction, it is beneficial if the orientation of the document is assessed correctly. Optionally, correct the orientation by dragging the rotation slider.

Optionally correct and press the "OK" button to continue. This will activate the automatic anonymization.

4. **Verify the anonymization, and optionally correct**

The user should verify whether the automatic anonymization is correct. When it is not correct, the user should improve the anonymization manually. Manual improvement is important for two reasons. First, manual improvement allows the user to modify the masks and guarantee that the exported images contain no personal data. Second, manual improvement is feedback for the anonymization module to learn and improve future automatic anonymizations. The user can improve the anonymization by using the visual elements in GUI. The GUI contains the following visual elements.

On the left:

- **Full image with overlays**: The full image is shown on the left. The user can interact with this image. This image contains two types of overlays:
  - **Keyword boxes**: Keyword boxes (e.g., for "Name") are colored green.
  - **Mask boxes**: Mask boxes (e.g., for "Smith") are colored red.
- **Radio buttons**: The radio buttons on the left can be used to select the interaction with the overlay boxes in the full image:
  - **Select**: No interaction selected.
  - **Add:** Add new boxes. The "Types" are explained in the description of the center part of the GUI.
  - **Edit**: Modify the location of the existing boxes.
  - **Remove**: Remove one of the existing boxes.

Modify the location of the boxes, when needed, and press "Refresh/Save" to proceed.

In the center:

- **Type**: Select the type:
  - **Keyword**: A keyword without a mask (e.g., for the optional field "remarks", where the keyword is present but the field is empty).
  - **Mask**: A box for a mask without a keyword (e.g., a document number at the bottom of the document).
  - **Keyword + mask**: A pair of boxes for keyword and mask, where both are related (e.g., keyword = "Name", mask = "Smith").
  - **Number**: For numbers without a related keyword.
  - **Face**: For a photograph.
  - **Stamp**: For a stamp.
  - **Signature**: For a signature
  - **Barcode**: For a barcode.
- **Id**: The first document can be anonymized in random order. All subsequent documents should be anonymized in the same order, to recognize keywords and locations consistently. The "Id" helps to annotate in a consistent way.
- **Snippet**: The center-part of the screen shows a snippet for each keyword. This allows the user to perform a consistent annotation, even when the user is not able to translate the snippet. For example, if in the first document the snippet with ID=1 is related to the name, then also in the second document, the snippet with ID=1 should

correspond to the name. Snippets are extracted from the first document where they are created.
- **Label**: The label can be added in Latin characters besides the snippet. This label is not relevant for anonymization. (See: [D4FLY-D8.4]).
- **Type**: The type is not relevant for anonymization. (See: [D4FLY-D8.4])

On the right, in the "Anonymization" pane:

- **Detail selection**: The detail region shows the following elements:
  - **Keyword image**: Zoom-in image on the active keyword region.
  - **Mask image**: Zoom-in image on the active mask region.
- **Anonymized image**: The anonymized image is shown on the right. This is an image with masks burned into the image as black boxes.
- **Buttons**:
  - **Refresh/Save**: The anonymized image can be updated by pressing the button "Refresh/Save" (Below the full image with overlays.)
  - **Export**: When the user is satisfied with the anonymized image, this image can be exported by pushing the button "Export". (Below the anonymized image.)
  - **Next**: When the user loaded multiple images and the user is ready with the current image, then he/she can go to the next image by pressing the button "Next".

5. Export

When the user is satisfied with the anonymized image, this image can be exported as a PNG image file by pushing the button "Export".

6. Go to next scan

When the user loaded multiple images and the user is ready with the current image, then he/she can go to the next image by pressing the button "Next". The next image is automatically anonymized and shown on the screen and the user continues at step 3.

7. Retrain

A deep-learning model is used for automatic detection and recognition (e.g., country recognition, keyword detection). Retraining the model uses the corrections of the user to improve the performance of the model. The user can retrain the model by using the menu bar:

- Menu-bar → "Model" → "Retrain".

8. Continue or Stop

The user can load new scans (step 2) or stop the application. Use one of the following steps to close the application:

- Use the red cross in the top right of the window.
- Menu-bar → "File" → "Exit".