D8.1 Document anonymization module 1

Document Due Date: 29/02/2020
Document Submission Date: 27/02/2020

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 information, 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 anonymization of travel and breeder documents. The tool consists of a graphical user interface, document recognition, keyword recognition, 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 anonymization of 88% of the related personal data.

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	**D8.1**
Document Title	**Document anonymization modules 1**
Work Package reference	WP8 Document Verification
Delivery due date	29/02/2020 [M6]
Actual submission date	27/02/2020
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	24/01/2020	TNO	Initial draft by the authors.
0.2	31/01/2020	TNO	New version based on review from Ben Kromhout (IND) and Jeroen van Rest (TNO).
0.3	10/02/2020	TNO	New version based on review from Hans de Moel (RNM) and Alfonsas Juršėnas (BPTI).
0.4	20/02/2020	TNO	New version based on review from Zachary Goldberg (TRI/WP3) and Dimitris Kyriazanos (NCSRD/SAB)
1.0	24/02/2020	TNO	Final version based on review from Armin Reuter (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
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. These documents contain personal information, such as name, date of birth and national number. The personal information must be well protected and a data breach must be avoided at all times. One of the ways to protect the personal data is to minimize the sharing of personal data (see GDPR article 5-c)[1]. Anonymization removes the personal information (e.g., by replacing the personal information by a black bar) and can therefore be used to minimize the sharing of personal data (see GDPR article 25-1)[2].

Travel and breeder documents will be processed automatically to detect document 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. However, sharing of personal data should be minimized because the documents contain personal data. The 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. 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.

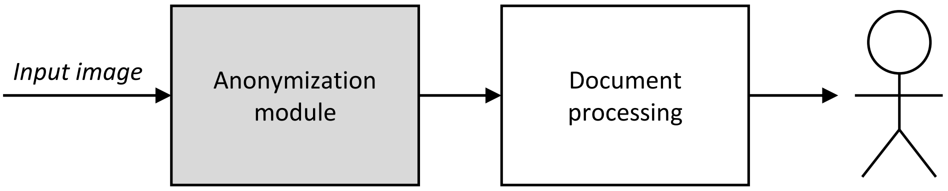


FIGURE 1-1: USE OF ANONYMIZATION BEFORE DOCUMENT PROCESSING.

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

[1] 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')".

[2] Art 25(1): "the controller shall (…) implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, 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."

whether the names are manipulated must be able to inspect those regions. In this case it is necessary to share the document including the personal information 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).

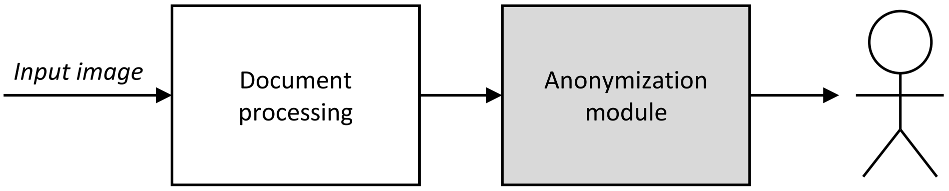


FIGURE 1-2: USE OF ANONYMIZATION AFTER DOCUMENT PROCESSING.

1.2 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 information in the scanned image and replace the personal information 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**: Start the application.
2. **Load data**: Load a set of images. The first image is automatically anonymized and shown on the screen.
3. **Improve manually**: The user can improve the anonymization manually, by creating new boxes or modifying/removing existing boxes. After that, the user confirms that the anonymization is verified and the anonymized image is updated.
4. **Export (optional):** The user can export the anonymized image.
5. **Next image (optional)**: The user can go to the next image. The next image is automatically anonymized and shown on the screen and the user continues at step 3.
6. **Continue or stop**: The user can load new data (step 2) or stop the application.

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.

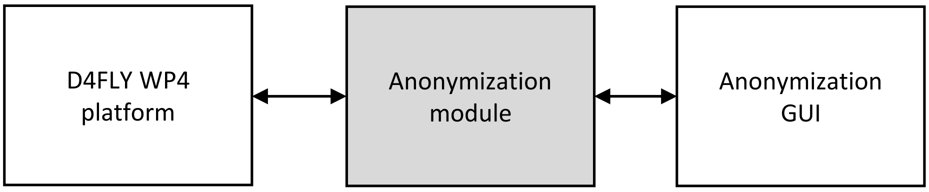


Figure 3-1: Top-level architecture.

3.2 Short interface description

There are three different interface messages between the anonymization GUI and the anonymization module. The normal workflow uses these interfaces in four steps (see Figure 3-2).

1. In the first step, an image is submitted from the GUI to the module. This image may contain personal information.
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 automatic anonymization message, which contains an update of the anonymized image. This anonymized image can be exported by the user.

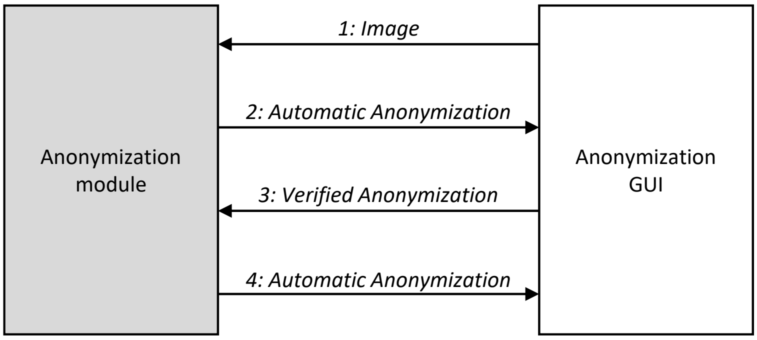

Anonymization module
1: Image
2: Automatic Anonymization
3: Verified Anonymization
4: Automatic Anonymization
Anonymization GUI

Figure 3-2: Interfaces.

Automatic batch processing can be implemented in the D4FLY platform by only using step 1 and 2. More details about the interface format can be found in Annex B.

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 information that is easy to read for machines.
- **Number without keyword**: The document may contain a (large) document number without any neighbouring 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.

4.1 Internal architecture

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

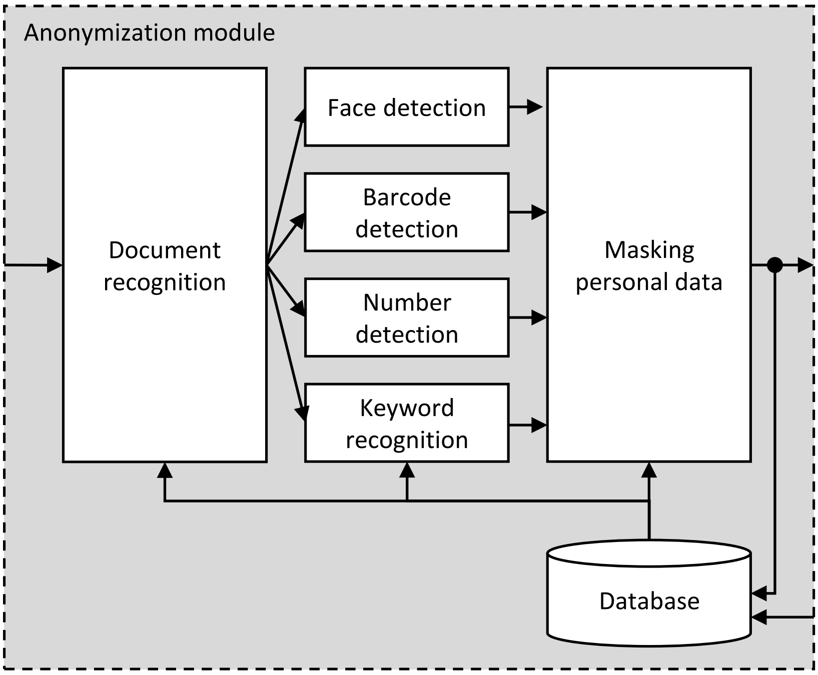


FIGURE 4-1: INTERNAL ARCHITECTURE.

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 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.

4.4 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 pretrained barcode detector PYZBAR [PYZBAR], which is a Python compatible open-source commonly-used zbar barcode detector.

4.5 Number detection

The module for “number detection” can localize numbers even when there are no keywords in the neighbourhood. 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.

4.6 Keyword recognition

Keywords (such as ‘First name’ or ‘Last name’ or ‘Date of birth’) can be detected with optical character recognition (OCR) or with a specific keyword detector – which detects the keyword as a whole. Both are explained in the following sections.

4.6.1 OCR

Optical Character Recognition (OCR) is implemented with TESSERACT [TESSERACT_OCR], which is an open source OCR engine. This implementation localizes words in the images. These words are subsequently classified as potential keywords using a keyword-dictionary. The image pre-processing consists of rotation, multiple color transformations, and application of OCR to each of the color transforms. The recognized text is merged and, for overlapping duplicates, the least confident are removed.

4.6.2 Keyword detection and localization

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.

4.6.3 Combine

The advantage of OCR is that it is pretrained and therefore it can immediately detect keywords in images of documents, which are presented to the software for the first time. The advantage of keyword detection is that it may reach better performance when many training examples are present. Therefore, we investigate the option of combining the results of the two keyword recognition modules with a 'union' operation. The ‘union’ operation keeps the boxes (big and small) and preserves them all which reduces the chance of missing keywords. This design decision depends on the accuracy of both detectors and will be explained in section 5.5.

4.7 Masking of textual data

The personal information 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 information) 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 verified document images. 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. Finally, when all relative distances are calculated, the 25th percentile (top-left coordinate of the mask) and 75th percentile (bottom-right coordinate of the mask) are used as a ‘template offset’ for that mask-keyword combination.

The second step applies these template offsets using the detected keywords. For every keyword that should be in the document, it is checked if the keyword is detected. Furthermore, when the detected keyword is not near the expected place in the document, it is treated as a missing keyword. If the keyword is detected, the direct keyword-mask offset is used. If a keyword is not detected, the keyword-mask offset from the nearest keyword is used.

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.

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 information 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 98%.

Table 5-1 Confusion matrix for document recognition.

	Country A (pred)	Country B (pred)	Country C (pred)
Country A (gt)	**20**	0	0
Country B (gt)	0	**19**	1
Country C (gt)	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 5-5, 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 was pretrained on an out-domain training set and validated on a small in-domain test set with 8 QR-codes. The results are shown in Table 5-3, where TP is true positive, FP is false positive and FN is false negative. The results show that almost 90% of the

QR-codes is correctly detected. The results seem very promising, but we should be careful with bold claims due to the small size of the evaluation set. Evaluation on a larger test set is planned as future work.

TABLE 5-3 RESULTS FOR BARCODE DETECTION.

	Number of barcodes	TP	FP	FN
QR codes	8	7	0	1

5.5 Keyword recognition

The keyword recognition was initially performed with keyword detection and OCR.

The pretrained OCR was applied to the dataset with 60 images. The advantage of OCR is that it does not need any annotated images for training. The results are shown in Table 5-4. The table shows that Latin alphabet is performing at almost 80%, but Non-Latin alphabets (including Arabic) are below 40%, which is probably not acceptable for end users.

TABLE 5-4 RESULTS FOR KEYWORD RECOGNITION WITH OCR.

Document type	OCR precision
Latin alphabet	78%
Non-Latin alphabet	26%

The keyword detection is trained on a part of the dataset and tested on another part of the dataset. 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 5-5. The results show that only 5 annotated images are needed to reach a performance of 92% accuracy.

TABLE 5-5 RESULTS FOR KEYWORD RECOGNITION WITH THE KEYWORD DETECTOR.

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

OCR appears to perform much worse than the keyword detector, even when the keyword detector is trained on only one image (75%). Therefore, the design of the system is modified such that we no longer use the OCR but solely the keyword detector.

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 5-6. 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 5-6 RESULTS FOR NON-LATIN NUMBER DETECTION.

	Number of Numbers (TP+FN)	TP	FP	FN
Document numbers	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: TPR = TP / (TP + FN)

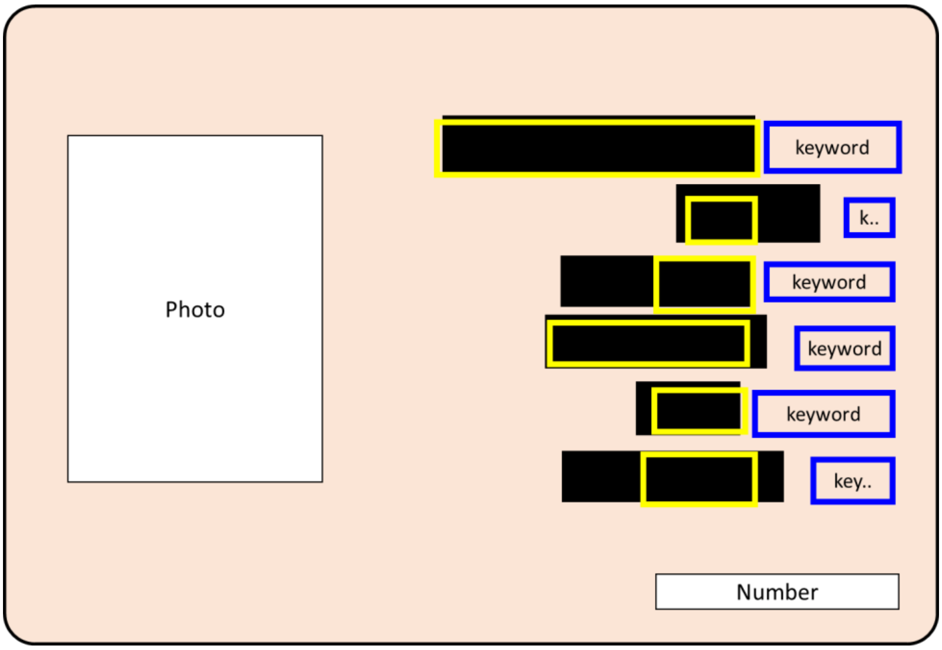
2) False Positive Rate: FPR = FP / (TP + FP),

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). The results are shown in Table 5-7. 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 TPR is 88%, which means that most masks are placed correctly. The average FPR is 41%, which indicates some over-segmentation. Figure 5-1 shows a representative example with an FPR of 40% 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).

TABLE 5-7 RESULTS FOR MASKING.

	Masking with ground-truth keyword boxes		Masking with automatic keyword detection	
	TPR	FPR	TPR	FPR
Country A	89%	39%	90%	38%
Country B	84%	44%	85%	44%
Country C	90%	44%	91%	43%
TOTAL	**88%**	**42%**	**88%**	**41%**

FIGURE 5-1: EXAMPLE RESULT OF MASKING TEXTUAL DATA WITH FPR=40%.



6 CONCLUSIONS AND FUTURE WORK

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 recognition, 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 anonymization of 88% of the related personal data. Face and barcode detection also reached a high accuracy of approximately 90%. Only the OCR-based number detection has lower performance but can be compensated by approximate localization of masks based on keywords.

Future work in D4FLY towards month M12 will include the following:

- **Retraining**: The current version can be retrained by scientists that developed the module. The next version will allow retraining by the end users.
- **Rotate**: The current version applies automatically rotation before OCR. The new version will include rotation before mask generation, and therefore also before the creation of the ground truth. This will help to align manual boxes in a more consistent way (and reduce the over-segmentation), and the automatic rotation can be corrected by the human.
- **Barcode**: The current version of barcode detection is tested on a very small dataset. The new version will be tested on a larger collection.
- **OCR**: Currently, OCR is performing poorly. For the anonymization module and for other modules, it will be important to improve OCR. Furthermore, OCR could improve the localization of the masks.
- **Usability**: Usability will be further investigated, e.g. to assess whether batch processing is desirable.

REFERENCES

[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).

[DLIB] http://dlib.net

[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).

[Hermans, 2017] Hermans, A., Beyer, L., & Leibe, B., “In defense of the triplet loss for person re-identification,” arXiv:1703.07737, (2017).

[PYZBAR] https://github.com/NaturalHistoryMuseum/pyzbar

[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

Use the following steps to start the application:

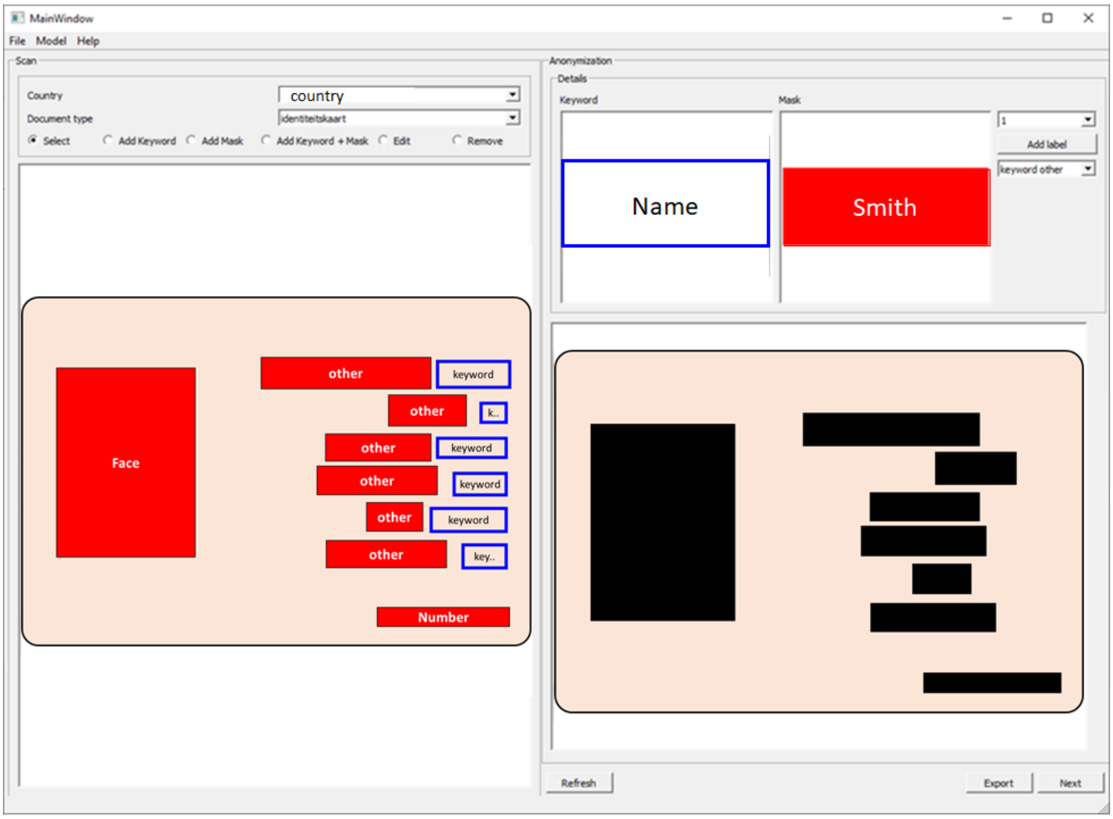
- Select on the desktop → 'Annotation Tool'.

2. Load data

Use the following steps to load data:

- Select on the top menu-bar → 'File' → 'Load data'.
- Browse to the folder with input images.
- Select one or multiple files.
- Select the 'Open' button.

The first image is automatically anonymized and 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.



Country: country
Document type: identiteitskaart
● Select ○ Add Keyword ○ Add Mask ○ Add Keyword + Mask ○ Edit ○ Remove
Keyword: Name
Mask: Smith
1
keyword other

3. Improve manually

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 information. 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 the GUI. The GUI contains the following visual elements.

- **Menu bar**: The menu bar (at the top) can be used to load data, exit the application or access help material.
- **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.
- **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 can be used to select the interaction with the overlay boxes in the full image:
    - **Select**: No interaction selected.
    - **Add keyword**: Add a new box for a keyword without a mask (e.g., for the optional field 'remarks', where the keyword is present but the field is empty).
    - **Add mask**: Add a new box for a mask without a keyword (e.g., a document number at the bottom of the document).
    - **Add keyword + mask**: Add a pair of boxes for keyword and mask, where both are related (e.g., keyword = "Name", mask = "Smith").
    - **Edit**: Modify the location of the existing boxes.
    - **Remove**: Remove one of the existing boxes.
- **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.
    - **Label dropdown**: Edit label index of the keyword.
    - **Add label button**: Add a new index to the list of indices in the label dropdown
    - **Keyword and mask type dropdown**: Meta data about the keyword (date, number, other.) or mask (face, barcode, number, other.)
- **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**: The anonymized image can be updated by pressing the button 'Refresh'.
    - **Export**: When the user is satisfied with the anonymized image, this image can be exported by pushing the button 'Export'.
    - **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'.

4. 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'.

5. Next image

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.

6. Continue or Stop

The user can load new data (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' → 'Close'.

Annex B: Interface format

The interface uses Protobuf-3 for serialization and ZeroMQ for data transfer. There are three main interface messages (as shown in the introduction to the interfaces in Chapter 3):

- Image
- Automatic_Anonymization
- Verified_Anonymization

Furthermore, there are several supporting messages:

- Anonymization
- Bbox
- Group

Each of the main messages and the supporting messages is defined below.

```
message Image {
  string id = 1; // Image UUID
  int32 width = 2; // Image width
  int32 height = 3; // Image height
  ImageType type = 4; // Image type
  bytes data = 5; // Image data

  enum ImageType {
    JPEG = 0;
    PNG = 1;
    TIFF = 2;
    BMP = 3;
  }
}
```

```
message Bbox {
  string id = 1; // Unique bbox UUID
  int32 x = 2; // X position of the bbox
  int32 y = 3; // Y position of the bbox
  int32 w = 4; // Width of the bbox
  int32 h = 5; // Height of the bbox
}
```

```
message Group {
  string id = 1; // Unique group UUID
  Bbox keyword_bbox = 2; // The bbox surrounding the keyword
  Bbox mask_bbox = 3; // The masking bbox to anonymize the personal information
  string keyword_label = 4; // Keyword description
  string keyword_label_id = 5; // The keyword UUID
  GroupType type = 6; // Group type
  string group_type = 7; // Group type description

  enum GroupType {
    FACE = 0;
    BARCODE = 1;
    NUMBER = 2;
    KEYWORD_DATE = 3;
    KEYWORD_NUMBER = 4;
    KEYWORD_OTHER = 5;
    MASK_OTHER = 6;
  }
}
```

```
message Anonymization {
  string id = 1; // Unique anonymization UUID
  string image_id = 2; // UUID of the image to anonymize
  string country_code = 3; // 2 letter country code (ISO 3166-1 alpha-2 code)
  string document_type_id = 4; // Document type UUID
  repeated Group groups = 5; // Groups with a keyword, mask and label
  Bbox document_title_bbox = 6; // Region indicating the document title
  string document_type = 7; // Document type (e.g., Birth certificate)
  string country = 8; // Country name.
}
```

```
message Automatic_Anonymization {
  Anonymization anonymization = 1;
  Image anonymized_image = 2; // The anonymized image
}
```

```
message Verified_Anonymization {
  Anonymization anonymization = 1;
}
```