UNIVERSITEIT•STELLENBOSCH•UNIVERSITY
jou kennisvennoot • your knowledge partner

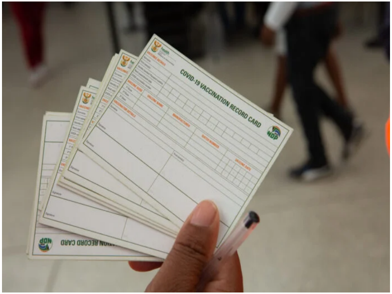
A framework for intelligent document image enhancement in pursuit of improved OCR performance

Ryno Kleinhans*
Supervisor: Dr GS Nel

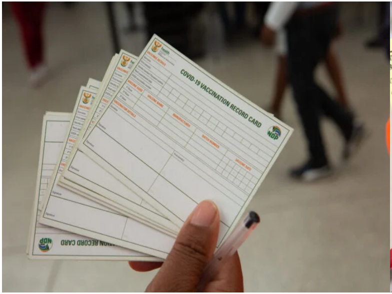
SUnORE

Stellenbosch Unit for Operations Research in Engineering
Department of Industrial Engineering

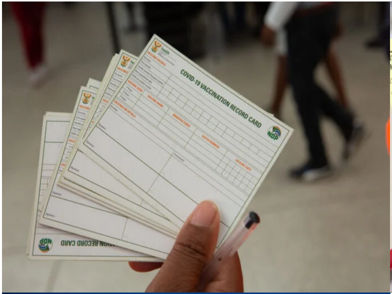
Overview of the problem - Paper-based documents

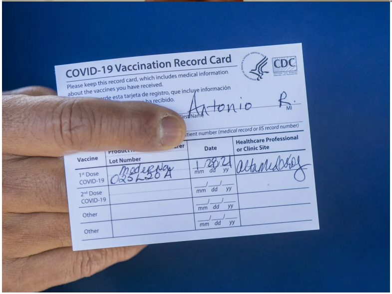


Overview of the problem - Paper-based documents

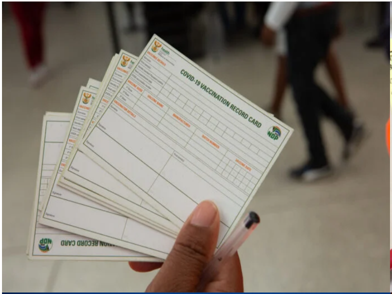


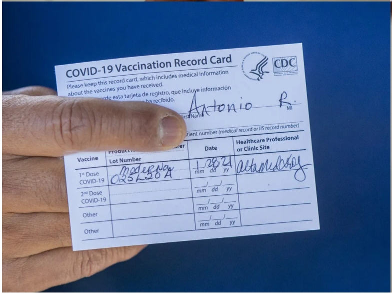
Overview of the problem - Paper-based documents

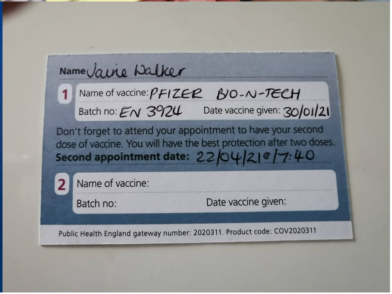




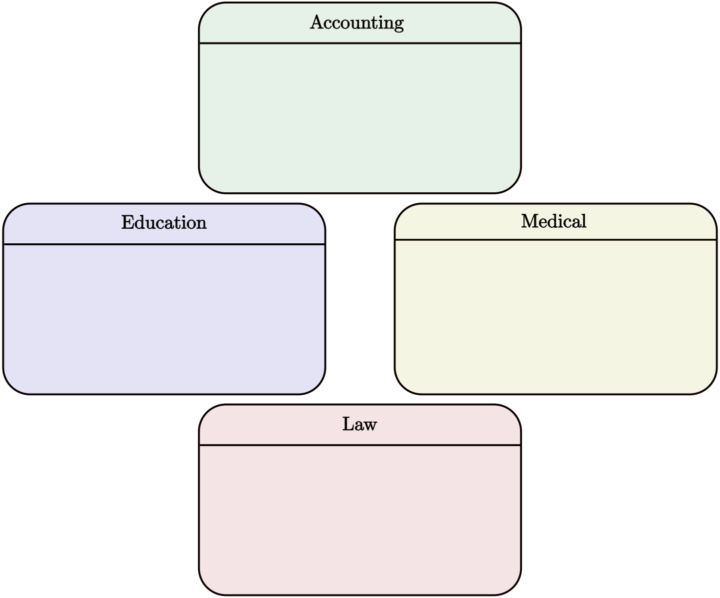
Overview of the problem - Paper-based documents







Overview of the problem - Paper-based documents



Overview of the problem - Paper-based documents

Accounting

- Single accountant prints 432 sheets of paper yearly
- 0.57 billion pages yearly (USA)
- Application forms, balance sheets, payslips, receipts

Education

Medical

Law

Overview of the problem - Paper-based documents

Accounting

- Single accountant prints 432 sheets of paper yearly
- 0.57 billion pages yearly (USA)
- Application forms, balance sheets, payslips, receipts

Education

- Single school prints 360 000 sheets of paper yearly
- 47 billion pages yearly (USA)
- Assignments, study materials, administrative documents

Medical

Law

Overview of the problem - Paper-based documents

Accounting

- Single accountant prints 432 sheets of paper yearly
- 0.57 billion pages yearly (USA)
- Application forms, balance sheets, payslips, receipts

Education

- Single school prints 360 000 sheets of paper yearly
- 47 billion pages yearly (USA)
- Assignments, study materials, administrative documents

Medical

- Single large hospital prints 96 million sheets of paper yearly
- 59 billion pages yearly (USA)
- Application forms, risk forms administrative documents

Law

Overview of the problem - Paper-based documents

Accounting

- Single accountant prints 432 sheets of paper yearly
- 0.57 billion pages yearly (USA)
- Application forms, balance sheets, payslips, receipts

Education

- Single school prints 360 000 sheets of paper yearly
- 47 billion pages yearly (USA)
- Assignments, study materials, administrative documents

Medical

- Single large hospital prints 96 million sheets of paper yearly
- 59 billion pages yearly (USA)
- Application forms, risk forms administrative documents

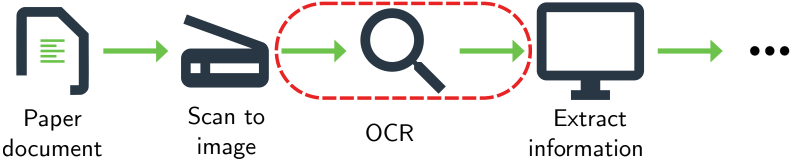
Law

- Single attorney prints 60 000 sheets of paper yearly
- 78 billion pages yearly (USA)
- Contracts, administrative documents, letters

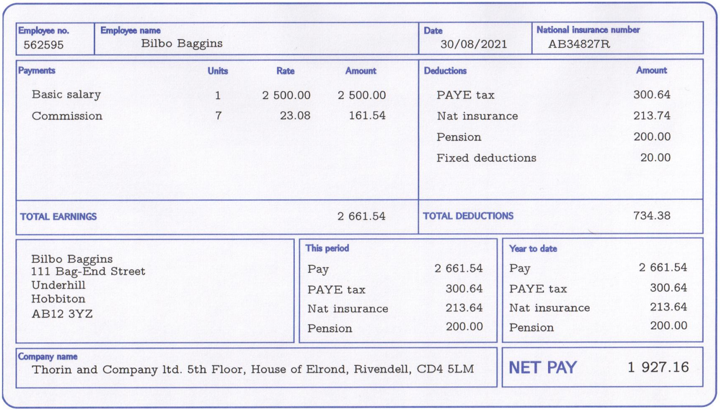
Overview of the problem - Automation example

OCR definition

Optical character recognition (OCR) is the electronic conversion of pixel-based text data (*i.e.* a captured image) comprising typed, handwritten or printed text into machine-encoded text.



Overview of the problem - Image processing



Employee no.	Employee name	Date	National insurance number
562595	Bilbo Baggins	30/08/2021	AB34827R

Payments	Units	Rate	Amount
Basic salary	1	2 500.00	2 500.00
Commission	7	23.08	161.54
TOTAL EARNINGS			2 661.54

Deductions	Amount
PAYE tax	300.64
Nat insurance	213.74
Pension	200.00
Fixed deductions	20.00
TOTAL DEDUCTIONS	734.38

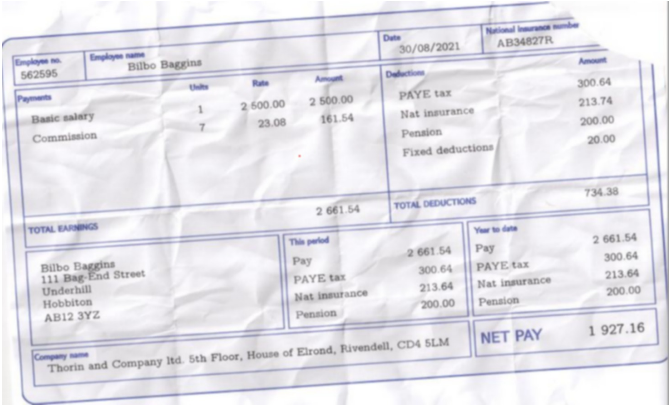
Bilbo Baggins
111 Bag-End Street
Underhill
Hobbiton
AB12 3YZ

This period	
Pay	2 661.54
PAYE tax	300.64
Nat insurance	213.64
Pension	200.00

Year to date	
Pay	2 661.54
PAYE tax	300.64
Nat insurance	213.64
Pension	200.00

Company name
Thorin and Company ltd. 5th Floor, House of Elrond, Rivendell, CD4 5LM

NET PAY 1 927.16



Employee no.
562595

Employee name
Bilbo Baggins

Date
30/08/2021

National Insurance number
AB34827R

Payments	Units	Rate	Amount
Basic salary	1	2 500.00	2 500.00
Commission	7	23.08	161.54
TOTAL EARNINGS			2 661.54

Deductions	Amount
PAYE tax	300.64
Nat insurance	213.74
Pension	200.00
Fixed deductions	20.00
TOTAL DEDUCTIONS	734.38

Bilbo Baggins
111 Bag-End Street
Underhill
Hobbiton
AB12 3YZ

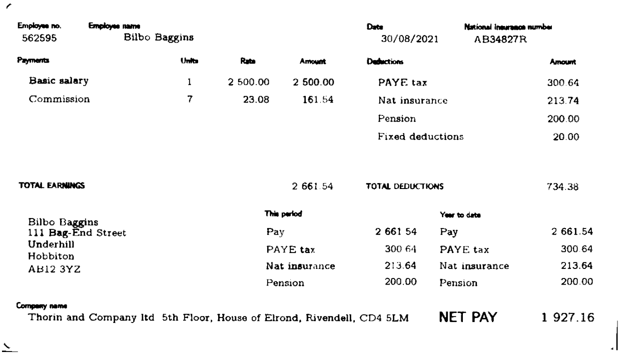
This period	
Pay	2 661.54
PAYE tax	300.64
Nat insurance	213.64
Pension	200.00

Year to date	
Pay	2 661.54
PAYE tax	300.64
Nat insurance	213.64
Pension	200.00

NET PAY 1 927.16

Company name
Thorin and Company ltd. 5th Floor, House of Elrond, Rivendell, CD4 5LM

Overview of the problem - Image processing



Employee no.	Employee name				Date	National insurance number	
562595	Bilbo Baggins				30/08/2021	AB34827R	

Payments	Units	Rate	Amount	Deductions	Amount
Basic salary	1	2 500.00	2 500.00	PAYE tax	300.64
Commission	7	23.08	161.54	Nat insurance	213.74
				Pension	200.00
				Fixed deductions	20.00
TOTAL EARNINGS			2 661.54	TOTAL DEDUCTIONS	734.38

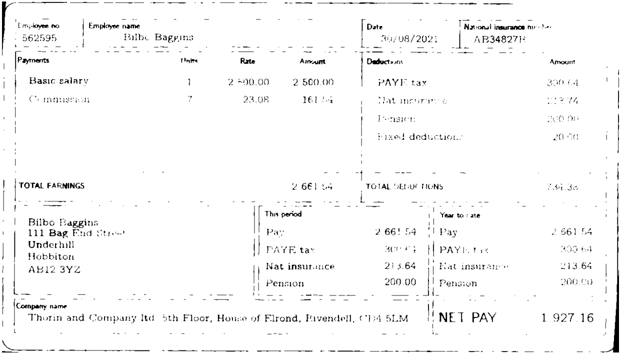
Bilbo Baggins
111 Bag-End Street
Underhill
Hobbiton
AB12 3YZ

This period		Year to date	
Pay	2 661 54	Pay	2 661.54
PAYE tax	300 64	PAYE tax	300.64
Nat insurance	213.64	Nat insurance	213.64
Pension	200.00	Pension	200.00

Company name
Thorin and Company ltd 5th Floor, House of Elrond, Rivendell, CD4 5LM

NET PAY 1 927.16

Overview of the problem - Image processing



Employee no	Employee name	Date	National insurance number
562595	Bilbo Baggins	30/08/2021	AB34827B

Payments	Units	Rate	Amount	Deductions	Amount
Basic salary	1	2 500.00	2 500.00	PAYE tax	300.64
Commission	7	23.08	161.54	Nat insurance	213.74
				Pension	200.00
				Fixed deductions	20.00
TOTAL EARNINGS			2 661.54	TOTAL DEDUCTIONS	734.38

Bilbo Baggins
111 Bag End Street
Underhill
Hobbiton
AB12 3YZ

This period		Year to date	
Pay	2 661.54	Pay	2 661.54
PAYE tax	300.64	PAYE tax	300.64
Nat insurance	213.64	Nat insurance	213.64
Pension	200.00	Pension	200.00

Company name
Thorin and Company ltd 5th Floor, House of Elrond, Rivendell, CB4 5LM

NET PAY 1 927.16

Agenda

Agenda

1. Informal problem description

Agenda

1. Informal problem description
2. Proposed framework walk-through

Agenda

1. Informal problem description
2. Proposed framework walk-through
3. Real-world payslip case study

Agenda

1. Informal problem description
2. Proposed framework walk-through
3. Real-world payslip case study
4. Receipt case study

Agenda

1. Informal problem description
2. Proposed framework walk-through
3. Real-world payslip case study
4. Receipt case study
5. Future work

Problem description

Informal problem description

The principal aim in this research project is to design, develop and demonstrate the practical workability of a **generic framework**

Problem description

Informal problem description

The principal aim in this research project is to design, develop and demonstrate the practical workability of a **generic framework** that introduces **machine intelligence** to the digitalisation of document images

Problem description

Informal problem description

The principal aim in this research project is to design, develop and demonstrate the practical workability of a **generic framework** that introduces **machine intelligence** to the digitalisation of document images in order to **improve OCR performance**.

Problem description

Informal problem description

The principal aim in this research project is to design, develop and demonstrate the practical workability of a **generic framework** that introduces **machine intelligence** to the digitalisation of document images in order to **improve OCR performance**.

- Available data:

Problem description

Informal problem description

The principal aim in this research project is to design, develop and demonstrate the practical workability of a **generic framework** that introduces **machine intelligence** to the digitalisation of document images in order to **improve OCR performance**.

- Available data:
  - Previously captured document images

Problem description

Informal problem description

The principal aim in this research project is to design, develop and demonstrate the practical workability of a **generic framework** that introduces **machine intelligence** to the digitalisation of document images in order to **improve OCR performance**.

- Available data:
  - Previously captured document images
  - Corresponding annotations (*i.e.* manually captured information)

Problem description

Informal problem description

The principal aim in this research project is to design, develop and demonstrate the practical workability of a **generic framework** that introduces **machine intelligence** to the digitalisation of document images in order to **improve OCR performance**.

- Available data:
  - Previously captured document images
  - Corresponding annotations (*i.e.* manually captured information)
- Tools and methods utilised in framework:

Problem description

Informal problem description

The principal aim in this research project is to design, develop and demonstrate the practical workability of a **generic framework** that introduces **machine intelligence** to the digitalisation of document images in order to **improve OCR performance**.

- Available data:
  - Previously captured document images
  - Corresponding annotations (*i.e.* manually captured information)
- Tools and methods utilised in framework:
  - Supervised learning paradigm

Problem description

Informal problem description

The principal aim in this research project is to design, develop and demonstrate the practical workability of a **generic framework** that introduces **machine intelligence** to the digitalisation of document images in order to **improve OCR performance**.

- Available data:
  - Previously captured document images
  - Corresponding annotations (*i.e.* manually captured information)
- Tools and methods utilised in framework:
  - Supervised learning paradigm
  - Pretrained convolutional neural networks

Problem description

Informal problem description

The principal aim in this research project is to design, develop and demonstrate the practical workability of a **generic framework** that introduces **machine intelligence** to the digitalisation of document images in order to **improve OCR performance**.

- Available data:
  - Previously captured document images
  - Corresponding annotations (*i.e.* manually captured information)
- Tools and methods utilised in framework:
  - Supervised learning paradigm
  - Pretrained convolutional neural networks
  - Pretrained OCR software

Problem description

Informal problem description

The principal aim in this research project is to design, develop and demonstrate the practical workability of a **generic framework** that introduces **machine intelligence** to the digitalisation of document images in order to **improve OCR performance**.

- Available data:
  - Previously captured document images
  - Corresponding annotations (*i.e.* manually captured information)
- Tools and methods utilised in framework:
  - Supervised learning paradigm
  - Pretrained convolutional neural networks
  - Pretrained OCR software
  - Common document image enhancement techniques

Proposed framework

The framework should facilitate:

Proposed framework

The framework should facilitate:

1. The **preparation** of previously annotated data and its document images for analysis,

Proposed framework

The framework should facilitate:

1. The **preparation** of previously annotated data and its document images for analysis,
2. the **engineering** and **labelling** of various unique enhancement procedures,

Proposed framework

The framework should facilitate:

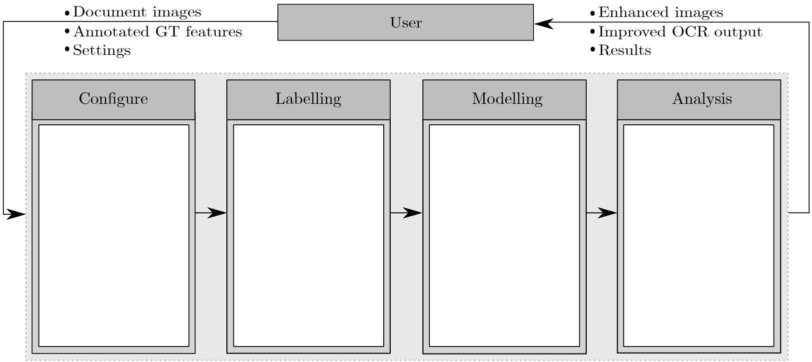
1. The **preparation** of previously annotated data and its document images for analysis,
2. the **engineering** and **labelling** of various unique enhancement procedures,
3. the **prediction** of the best enhancement procedure for each unique unseen document image,

Proposed framework

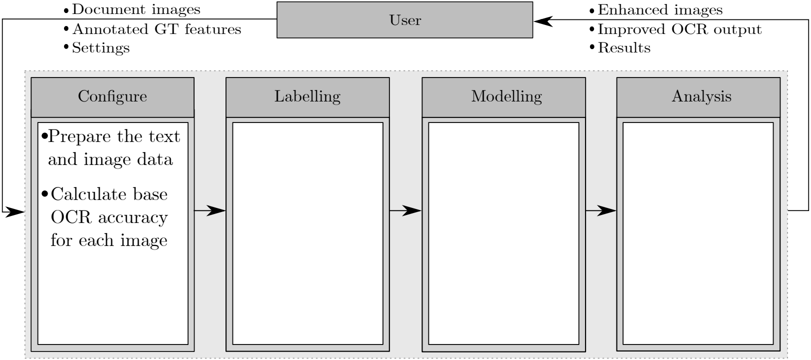
The framework should facilitate:

1. The **preparation** of previously annotated data and its document images for analysis,
2. the **engineering** and **labelling** of various unique enhancement procedures,
3. the **prediction** of the best enhancement procedure for each unique unseen document image,
4. as well as the **implementation** and **analysis** of the predictions.

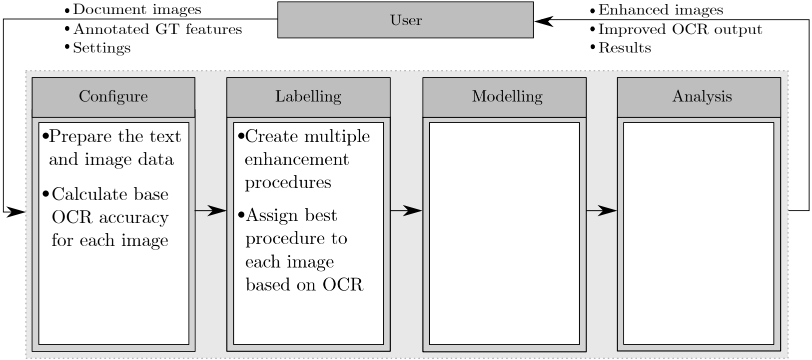
Proposed framework - Subcomponents



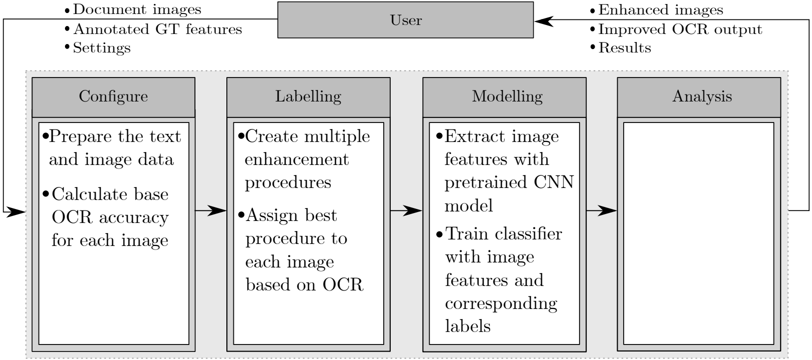
Proposed framework - Subcomponents



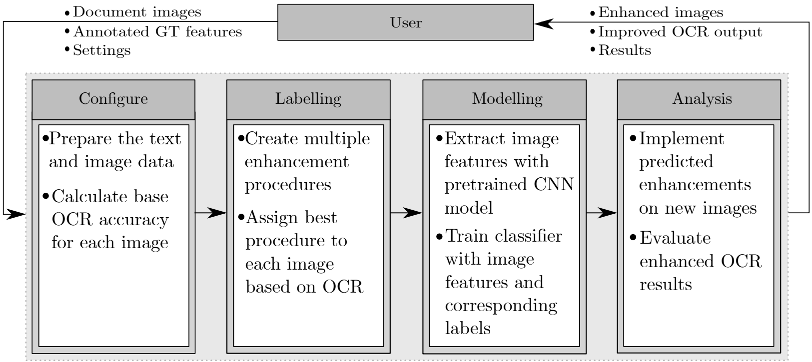
Proposed framework - Subcomponents



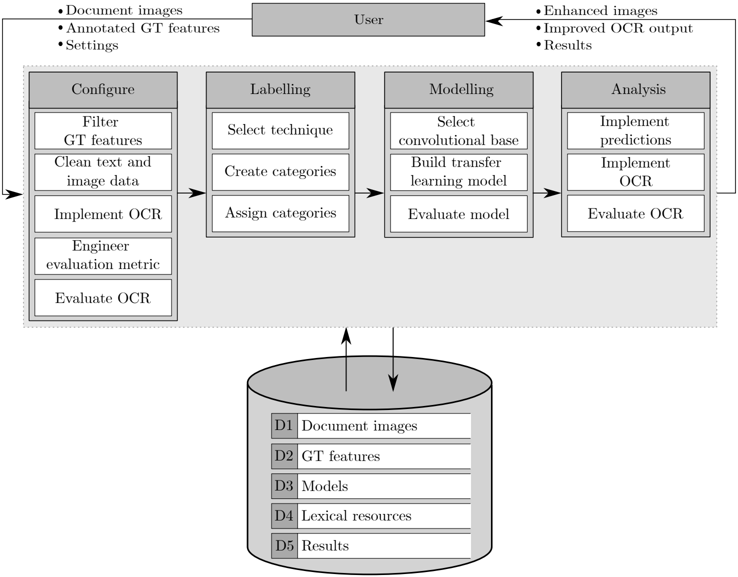
Proposed framework - Subcomponents



Proposed framework - Subcomponents



Proposed framework - Modules



Real-world case study - Industry partner data

- 2 000 PDFs of scanned client payslips

Real-world case study - Industry partner data

- 2 000 PDFs of scanned client payslips
- CSV with true values captured by branch consultant:

Real-world case study - Industry partner data

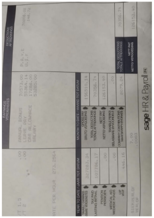
- 2 000 PDFs of scanned client payslips
- CSV with true values captured by branch consultant:
  - 2 000 rows of client information

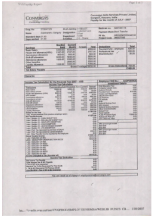
Real-world case study - Industry partner data

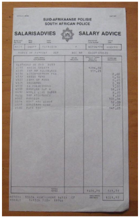
- 2 000 PDFs of scanned client payslips
- CSV with true values captured by branch consultant:
  - 2 000 rows of client information
  - 11 columns of captured features (*i.e.* Ground truth) comprising client name, net pay, total pay, and occupation.

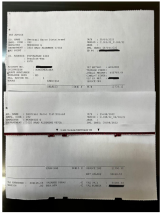
Real-world case study - Industry partner data

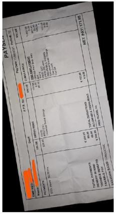
- 2 000 PDFs of scanned client payslips
- CSV with true values captured by branch consultant:
  - 2 000 rows of client information
  - 11 columns of captured features (*i.e.* Ground truth) comprising client name, net pay, total pay, and occupation.



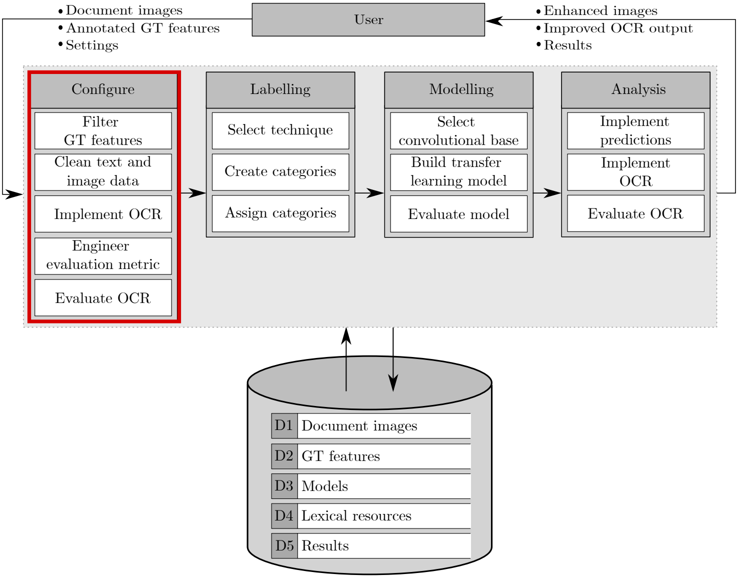




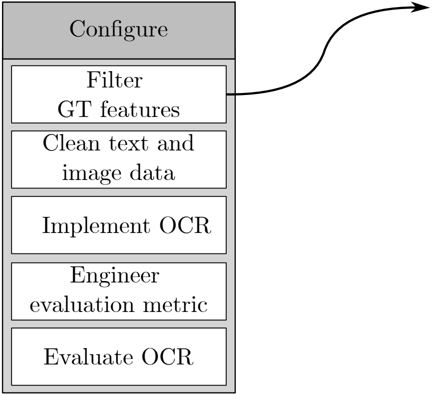




Real-world case study - Configure subcomponent

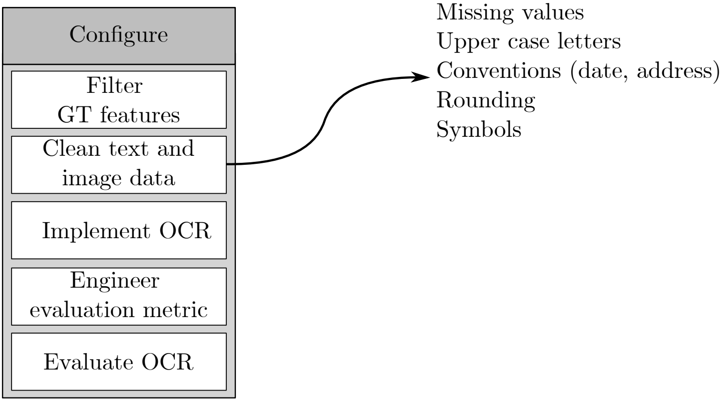


Real-world case study - Configure subcomponent

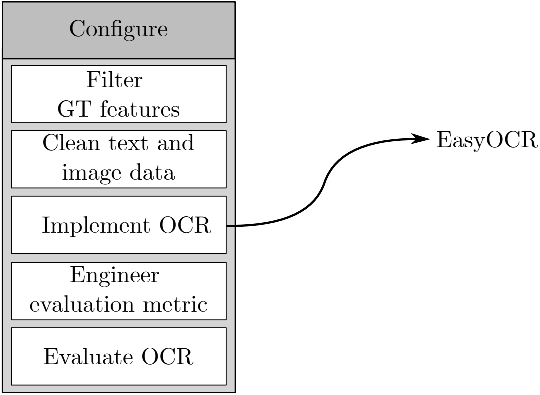


Employee code
Employee name
Employee surname
Date
Total earnings
Basic salary
Total deductions
Net pay
Company name
Occupation

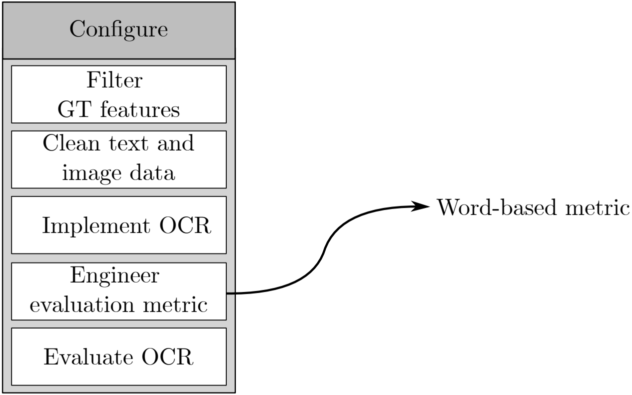
Real-world case study - Configure subcomponent



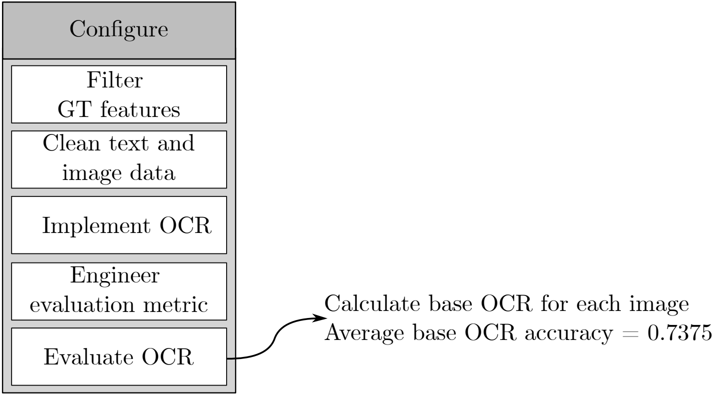
Real-world case study - Configure subcomponent



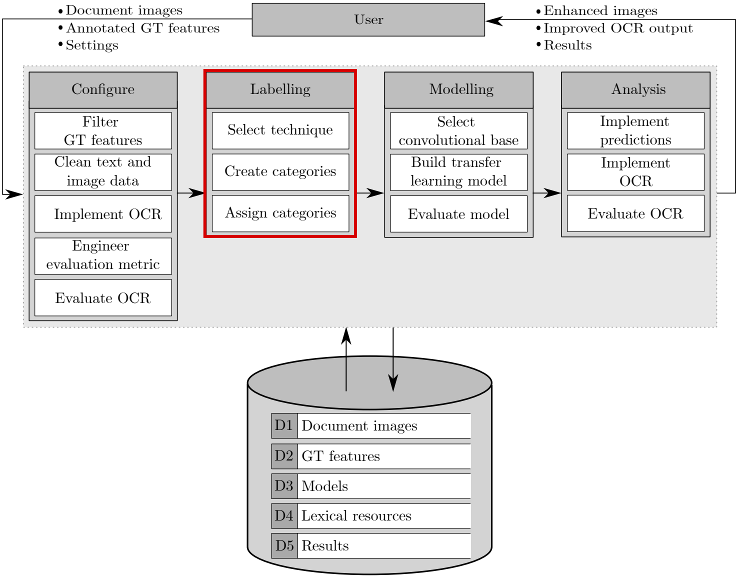
Real-world case study - Configure subcomponent



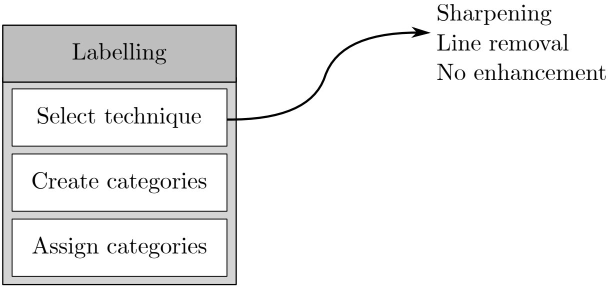
Real-world case study - Configure subcomponent



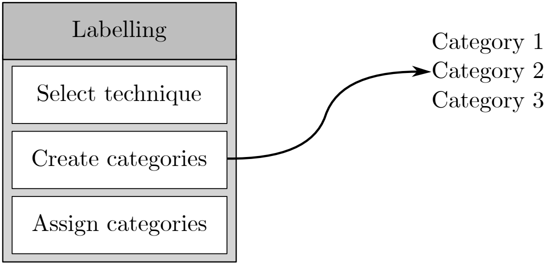
Real-world case study - Labelling subcomponent



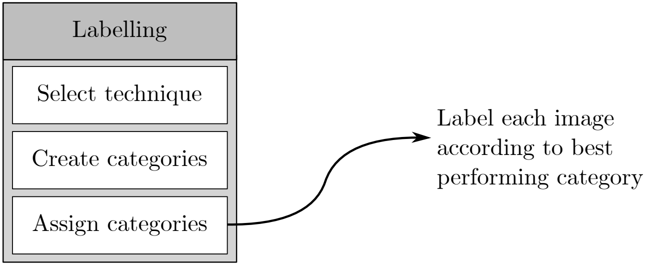
Real-world case study - Labelling subcomponent



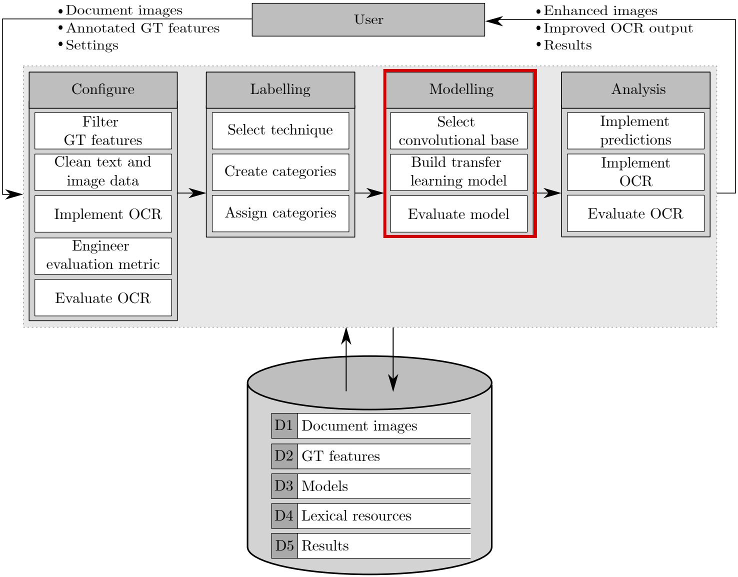
Real-world case study - Labelling subcomponent



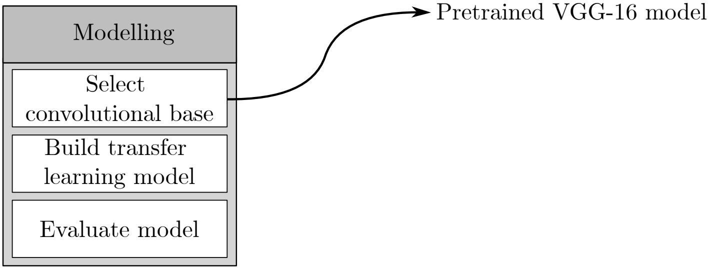
Real-world case study - Labelling subcomponent



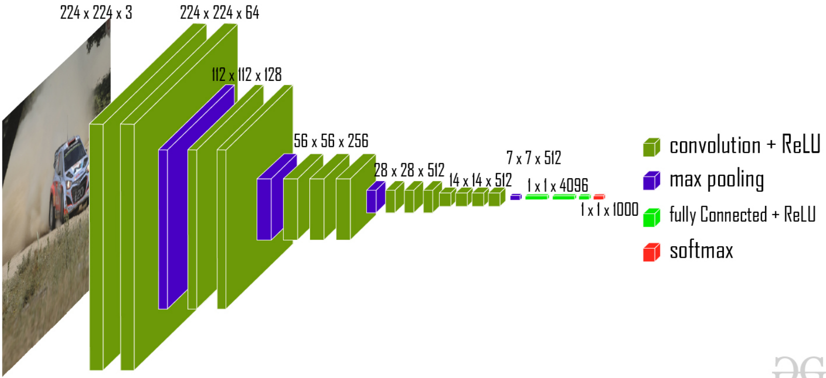
Real-world case study - Modelling subcomponent



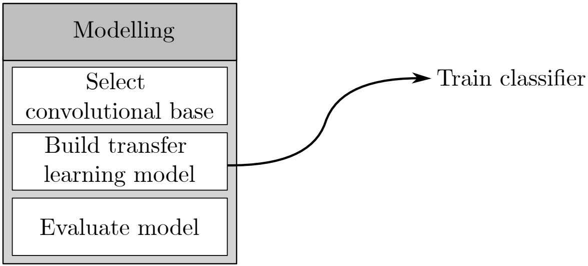
Real-world case study - Modelling subcomponent



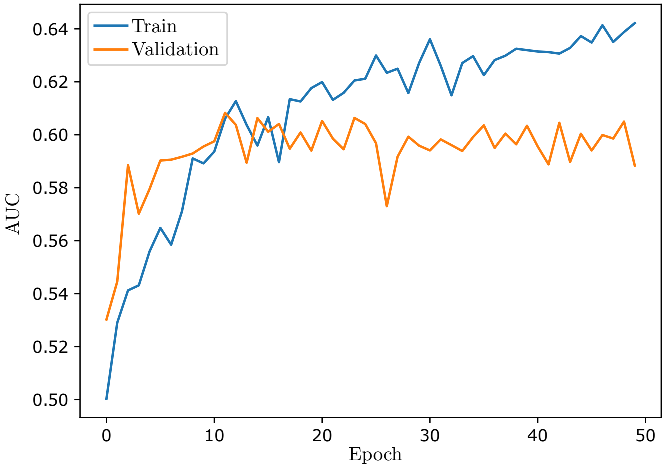
Real-world case study - Modelling subcomponent



Real-world case study - Modelling subcomponent



Real-world case study - Modelling subcomponent



Real-world case study - Modelling subcomponent

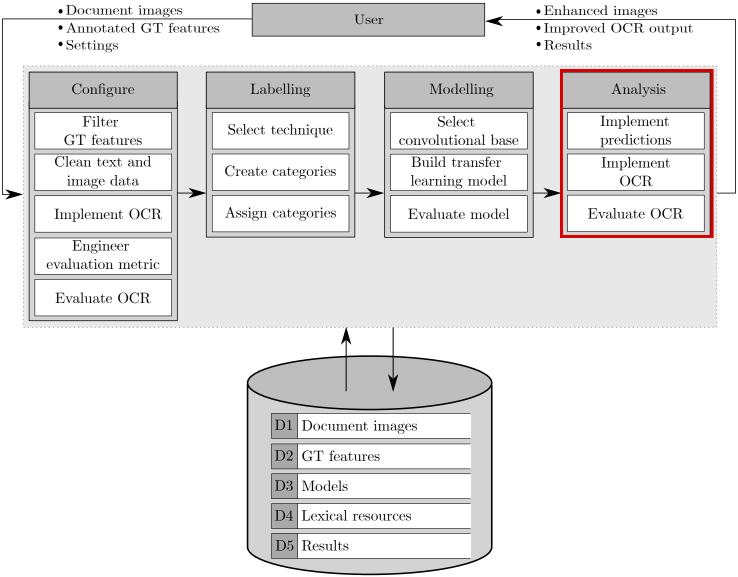
Predicted \ Actual	A	B	C
A	0.57	0.16	0.42
B	0.16	0.52	0.14
C	0.28	0.32	0.44

A: Base image

B: Line Removal

C: Sharpening

Real-world case study - Analysis subcomponent



Real-world case study - Analysis subcomponent

Test set category	Average OCR accuracy
Base	0.7382
Framework	0.7444

Real-world case study - Analysis subcomponent

Impact category compared with the original base OCR	Line removal on all images		Sharpening on all images		Only predicted images	
	Number	Ratio	Number	Ratio	Number	Ratio
Improved	60	0.16	73	0.20	59	0.16
Same	249	0.68	187	0.51	268	0.73
Deteriorated	59	0.16	108	0.30	41	0.11
Improved/deteriorated		**1.0169**		**0.6759**		**1.4390**

Real-world case study - Analysis subcomponent

Impact category compared with the original base OCR	Line removal on all images		Sharpening on all images		Only predicted images	
	Number	Ratio	Number	Ratio	Number	Ratio
Improved	60	0.16	73	0.20	59	0.16
Same	249	0.68	187	0.51	268	0.73
Deteriorated	59	0.16	108	0.30	41	0.11
Improved/deteriorated		**1.0169**		**0.6759**		**1.4390**

Real-world case study - Analysis subcomponent

Impact category compared with the original base OCR	Line removal on all images		Sharpening on all images		Only predicted images	
	Number	Ratio	Number	Ratio	Number	Ratio
Improved	60	0.16	73	0.20	59	0.16
Same	249	0.68	187	0.51	268	0.73
Deteriorated	59	0.16	108	0.30	41	0.11
Improved/deteriorated		**1.0169**		**0.6759**		**1.4390**

Real-world case study - Analysis subcomponent

Impact category compared with the original base OCR	Line removal on all images		Sharpening on all images		Only predicted images	
	Number	Ratio	Number	Ratio	Number	Ratio
Improved	60	0.16	73	0.20	59	0.16
Same	249	0.68	187	0.51	268	0.73
Deteriorated	59	0.16	108	0.30	41	0.11
Improved/deteriorated		**1.0169**		**0.6759**		**1.4390**

Case study 2 - Data provided by ICDAR

- 1 000 PDFs of scanned restaurant receipts

Case study 2 - Data provided by ICDAR

- 1 000 PDFs of scanned restaurant receipts
- CSV with true values captured by annotators:

Case study 2 - Data provided by ICDAR

- 1 000 PDFs of scanned restaurant receipts
- CSV with true values captured by annotators:
  - 1 000 rows of client information

Case study 2 - Data provided by ICDAR

- 1 000 PDFs of scanned restaurant receipts
- CSV with true values captured by annotators:
  - 1 000 rows of client information
  - 4 columns of captured features (*i.e.* Ground truth) comprising company name, date, address, and receipt total.

- 1 000 PDFs of scanned restaurant receipts
- CSV with true values captured by annotators:
  - 1 000 rows of client information
  - 4 columns of captured features (*i.e.* Ground truth) comprising company name, date, address, and receipt total.





Case study 2 - Receipt results

Case study 2 - Receipt results

Predicted \ Actual	A	B
A	0.61	0.43
B	0.39	0.57

A: Base image

B: Sharpening

Case study 2 - Receipt results

Test set category	Average OCR accuracy
Base	0.7720
Framework	0.7827

Case study 2 - Receipt results

Impact category compared with the original base OCR	All images in test set		Only applied to predicted images in test set	
	Number	Ratio	Number	Ratio
Improved	38	0.30	22	0.18
Same	39	0.31	83	0.66
Deteriorated	48	0.38	20	0.16
Improved/deteriorated		**0.7895**		**1.1250**

Case study 2 - Receipt results

Impact category compared with the original base OCR	All images in test set		Only applied to predicted images in test set	
	Number	Ratio	Number	Ratio
Improved	38	0.30	22	0.18
Same	39	0.31	83	0.66
Deteriorated	48	0.38	20	0.16
Improved/deteriorated		**0.7895**		**1.1250**

Case study 2 - Receipt results

Impact category compared with the original base OCR	All images in test set		Only applied to predicted images in test set	
	Number	Ratio	Number	Ratio
Improved	38	0.30	22	0.18
Same	39	0.31	83	0.66
Deteriorated	48	0.38	20	0.16
Improved/deteriorated		**0.7895**		**1.1250**

Future work

- Possible follow-up work on the contributions of this research project:

Future work

- Possible follow-up work on the contributions of this research project:
    1. Consider the inclusion of a confidence score for the OCR evaluation metric,

Future work

- Possible follow-up work on the contributions of this research project:
  1. Consider the inclusion of a confidence score for the OCR evaluation metric,
  2. investigate which pre-trained convolutional base filters are most important for extracting the intrinsic patterns within the provided data sets,

Future work

- Possible follow-up work on the contributions of this research project:
  1. Consider the inclusion of a confidence score for the OCR evaluation metric,
  2. investigate which pre-trained convolutional base filters are most important for extracting the intrinsic patterns within the provided data sets,
  3. applying the framework to document images with handwritten characters.

References

SMITH R, 2007, *An overview of the Tesseract OCR engine*, Ninth international conference on document analysis and recognition (ICDAR 2007), **2**, pp. 629–633.

Jaided AI, 2020, *EasyOCR*, [Online], [Cited September 2021], Available from https://github.com/JaidedAI/EasyOCR