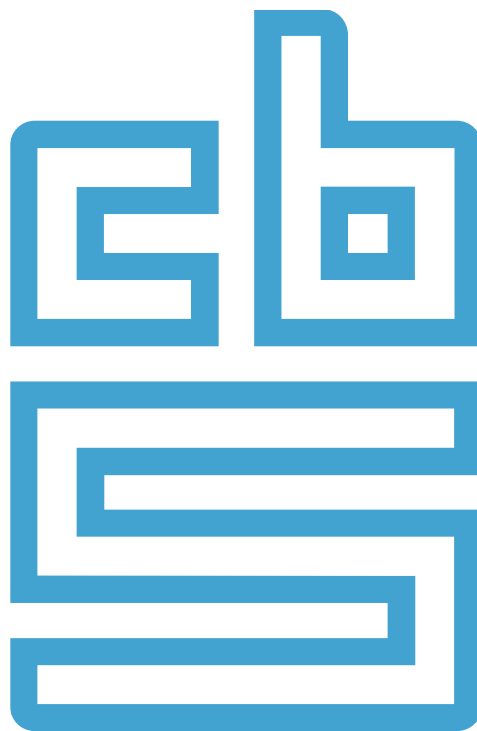






Smart surveys: active vs. passive data collection

A shop receipt scanning and classification pipeline for HBS

Nick de Wolf & Barry Schouten

MASS Workshop, 23-04-2021

Active – passive data collection

Smart surveys have to choose a level of active participation of respondents. This leads to a trade-off:

- Respondent burden: Passive
- Data quality: Active
- Respondent involvement: Active/passive on demand

Case study: Smart features in the Household Budget Survey:

- Receipt scanning and processing
- Donation of bank transactions data
- Linkage of scanner transactions data
- Geo-fencing



Respondent in-app interaction

What interactions may exist?

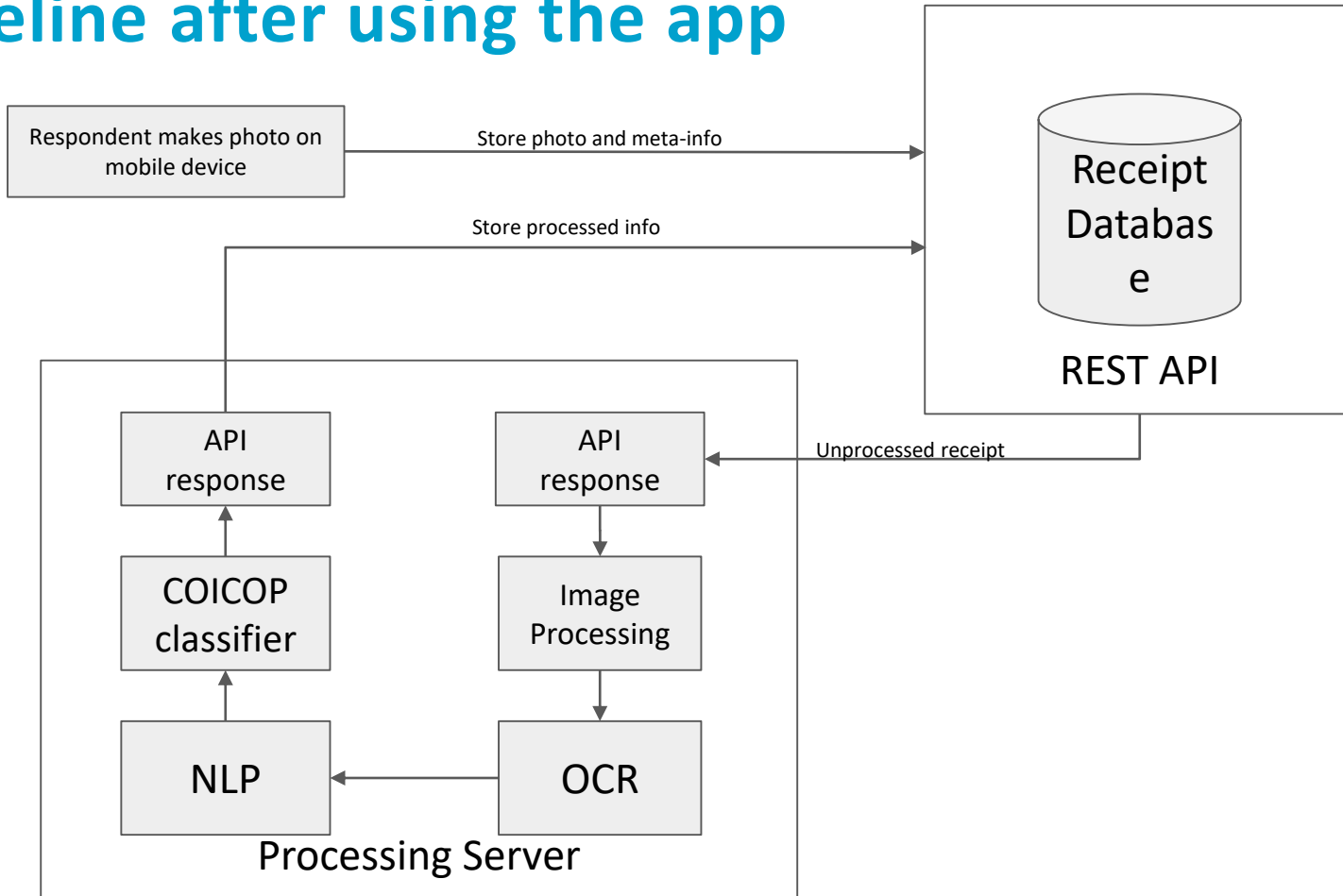
- Soft warnings for low quality scanning:
 - Lighting conditions
 - Visible contrast
- Bounding box for background removal
- Preliminary result of in-app OCR and NLP
- Validation of results of OCR and NLP
- Validation of results of product classification



In-app OCR and NLP demo



Pipeline after using the app



Backend data

- Image data and meta-data is sent to CBS database
 - Both raw image and preprocessed image are sent.
 - Metadata: such as shop, total price, country, online/physical purchase
- Image information is used for in-house OCR and COICOP classification
- Metadata can be used to verify classification results
- Results are returned to the respondent
 - Thresholds of certainty can be set before returning

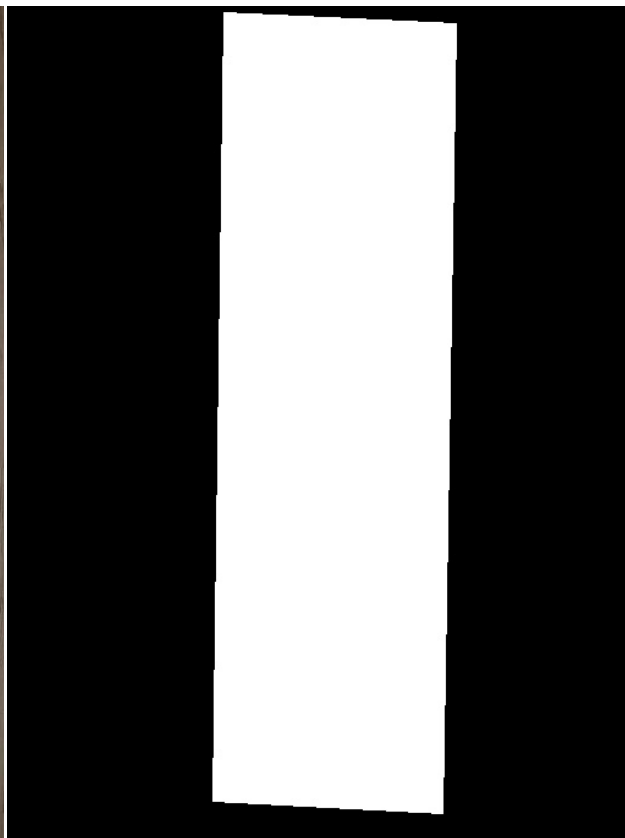


JUMBO

supermarkten

Smits van Oyenlaan 26
6076 DP Houten
0495565510
www.jumbo.com

OPSCHEFFING	BEDRAG IN €	
JUMBO KERSTAS 2019	0,35	
JUMBO MESCLUN MELANGE	1,69	
JUMBO KOKKOMMER	0,85	
PETERSSELIE	1,06	
JUMBO PEPER ROOD	1,19	
JUMBO ALFAFA	2,39	
JUMBO KORIANDER		
2 x 1,38	2,76	
JUMBO GETIENK BOLLETJE	1,75	
JUMBO RODE WIEN	0,66	
JUMBO BIO KNOFLOOK	0,69	
BIO SNOEPTOMAATJES	1,50	
KESBENE AIGURKBLOKJE	1,15	
AVOCADO	1,29	
Totaal	17,35	
Betaald:	17,35	
Kopie Kaarthouder		
Terminal: 0K2607	Merchant: 145896	
Periode: 9355	Transactie: 02370872	
Contactloze betaling		
MAESTRO (A0000000043060)		
MAESTRO		
Kaart: xxxxxxxxxxxxxxx9056		
Kaartnr: 00		
BETALING		
21/12/2019 16:50 Auth. code: 583089		
Totaal: 17,35 EUR		
AKKOORD		
	BTW overzicht	
BTW	Bedrag excl	BTW bedrag
9,00%	15,60	1,40
21,00%	0,29	0,06
BTW Totaal	15,89	1,46
Openingsijden:		
Maandag-Zaterdag	08:00-20:00	
Zondag	10:00-18:00	
WEEKDAG: 11:00		
TOTAAL: 3076 PPS, 3 TANKS 2005 15:50 21-12-2019		



Example



Part 1

WF SE LA
tit is 'aa Th ee Wa AU ghd
La if A LL Qu dk!

UPS) RP ak bees
Kuen) &

Smits van Uytenlaan 2b
6026 CP Haarheeze
0495590510
Wii, Jumbo. COM

ONSCHR VIN | BEDRAG IN 4 €
JUNBO "KERSTTAS. 2019 Dado
JMB MESCLUN MELANGE 1,65
| JUMBO KOMKOMMER 0,85
PETERSELIE 1,08
JUMBO PEPER ROOD 1,19
JUMBO ALFALFA 2 39
JUMBO KORTANDER |
JMB GEITENK BOLLETJE 1,09
JUMBO RODE UIEN 0,66
| JUMBO BIO KNOFLOOK 0,69
| BIO SNOEPTOMAATJES 1,50
| KESBEKE AUGURKBLOKJE 1,15
Totaal 17,35
| Betaald:
| Maestro 17,32

Kopie Kaarhouder

Terminal: QK2607 Merchant: 145896
Periode: 9355 Transactie: 02370872
Contactloze betaling
MAESTRO (A0000000043060)
MAESTRO
Kaart: xxxxxxxxxxxxxxx9056
Kaartnr: 00

Part 2

MAESTRO (A0000000043060)

MAESTRO

Kaart: XXXXXXXXXXXXXXX91056

Kaartnr: 00

BETALING

21/12/2019 16:50 Auth. code: 583089

Totaal: 17,35 EUR

— AKKOORD —

oe ne ee 'a
ee vo) 2 = FS ee wy Cal Lam uy PAS (ee (us (en) ay) ee a

ar ' arr ion

Bris Bedrag exc] BTW bedra
9 00% 15,60 140
21, 00% 0,29 i 06
BIW Totaal 15,9 1,46
| ' Openinget iden: i
Maandag-/aterdag 06 :00-20:00
/ fici JO :00 15 :U0

He VEE HIER Line
STORE. Shh Pus J TRANS 1005 16:50 21-12-2019

ae etre NA

636760033065127 127019

Worse quality image,
introduces more mistakes:

This examples happens to do the products
well, but other text shows errors such as:

- Jumbo -> Junbo
- B0,35 -> Dado
- 1,69 -> 1,65



Example

Part 1

WF SE LA
tit is 'aa Th ee Wa AU ghd
La if A LL Qu dk!

UPS] RP ak bees
Kuen) &

Smits van Uyentaan 2b
6026 CP Haarheeze
0495590510
Wii, Jumbo. COM

ONSCHR VIN | BEDRAG IN 4 €
JUNBO "KERSTTAS. 2019 Dado
JMB MESCLUN MELANGE 1,65
| JUMBO KOMKOMMER 0,85
PETERSELIE 1,08
JUMBG PEPPER ROOD 1,19
JUMBO ALFALFA 2 39
JUMBO KORTANDER |
JMB GEITENK BOLLETJE 1,09
JUMBO RODE UIEN 0,66
| JUMBO BIO KNOFLOOK 0,69
| BIO SNOEPTOMAATJES 1,50
:| KESBEKE AUGURKBLOKJE 1,15
| AVOCADO 1,29
Totaal 17,35
Betaald:
Maestro 17,32

Kopie Kaarhouder

Terminal: QK2607 Merchant: 145896
Periode: 9355 Transactie: 02370872
Contactloze betaling

Part 2

MAESTRO (A000000040360)

MAESTRO

Kaart: XXXXXXXXXXXXXXX91056

Kaartnr: OU

BETALING

21/12/2019 16:50 Auth. code: 583089
Totaal: 17,35 EUR
— AKKOORD —

oe ne ee ' a
ee vo) 2 = FS ee wy Cal Lam uy PAS (ee (us (en) ay) ee à

ar ' arr ion

Bris Bedrag exc] BTW bedra
9 00% 15,60 140
21 , 00% 0,29 i 06
BIW Totaal 15, 9 1,46
| ' Openinget iden: i
Maandag-/aterdag 06 :00.-2(:0(/
fic i JO : 00 15 :U0

He VEE HIER Line
STORE. Shh Pus J TRANS 1005 16:50 21-12-2019

ae etre NA

636760033065127 127019

	RECDISC	Price	Amount	Pay- mode	Shopname	Date
0	junbo kersttas. 2019 dado	0.0	1	card	Jumbo	21/12/2019
1	jmb mesclun melange	1.65	1	card	Jumbo	21/12/2019
2	jumbo komkommer	0.85	1	card	Jumbo	21/12/2019
3	peterselie	1.08	1	card	Jumbo	21/12/2019
4	jumbg peper rood	1.19	1	card	Jumbo	21/12/2019
5	jumbo alfalfa	2 39	1	card	Jumbo	21/12/2019
6	jumbo kortander	0.0	1	card	Jumbo	21/12/2019
7	jmb geitenk bolletje	1.09	1	card	Jumbo	21/12/2019
8	jumbo rode uien	0.66	1	card	Jumbo	21/12/2019
9	jumbo bio knoflook	0.69	1	card	Jumbo	21/12/2019
10	bio snoeptomaatjes	1.50	1	card	Jumbo	21/12/2019
11	: kesbeke augurkblokje	1.15	1	card	Jumbo	21/12/2019
12	avocado	1.29	1	card	Jumbo	21/12/2019



Example

	RECDESC	Price	Amount	Pay-mode	Shopname	Date	Coicop_class	Coicop_class_name	Coicop_class_prob
0	junbo kersttas. 2019 dado	0.0	1	card	Jumbo	21/12/2019	09.3.1.2	Speelgoed en feestartikelen	90
1	jmb mesclun melange	1.65	1	card	Jumbo	21/12/2019	01.1.7.1	Verse of gekoelde groenten behalve aardappelen...	100
2	jumbo komkommer	0.85	1	card	Jumbo	21/12/2019	01.1.7.1	Verse of gekoelde groenten behalve aardappelen...	64
3	petersellie	1.08	1	card	Jumbo	21/12/2019	01.1.9.2	Zout, specerijen en keukenkruiden	84
4	jumbg peper rood	1.19	1	card	Jumbo	21/12/2019	01.1.7.1	Verse of gekoelde groenten behalve aardappelen...	74
5	jumbo alfalfa	2.39	1	card	Jumbo	21/12/2019	01.1.7.1	Verse of gekoelde groenten behalve aardappelen...	99
6	jumbo kortander	0.0	1	card	Jumbo	21/12/2019	01.1.1.3	Brood	75
7	jmb geitenk bolletje	1.09	1	card	Jumbo	21/12/2019	01.1.8.3	Chocolade	96
8	jumbo rode uien	0.66	1	card	Jumbo	21/12/2019	01.1.6.1	Vers of gekoeld fruit	74
9	jumbo bio knofflook	0.69	1	card	Jumbo	21/12/2019	01.1.7.1	Verse of gekoelde groenten behalve aardappelen...	90
10	bio snoeptomaatjes	1.50	1	card	Jumbo	21/12/2019	01.1.7.1	Verse of gekoelde groenten behalve aardappelen...	100
11	: kesbeke augurkblokje	1.15	1	card	Jumbo	21/12/2019	01.1.7.3	Gedroogde groenten, andere bereidingen en cons...	100
12	avocado	1.29	1	card	Jumbo	21/12/2019	01.1.6.1	Vers of gekoeld fruit	82



Discussion

- What processing performance level is acceptable for respondent validation?
- Should we apply soft or hard warnings?
- Manual in-house verification: How frequent, if at all?
- Any experience with respondent validation?





Facts that matter