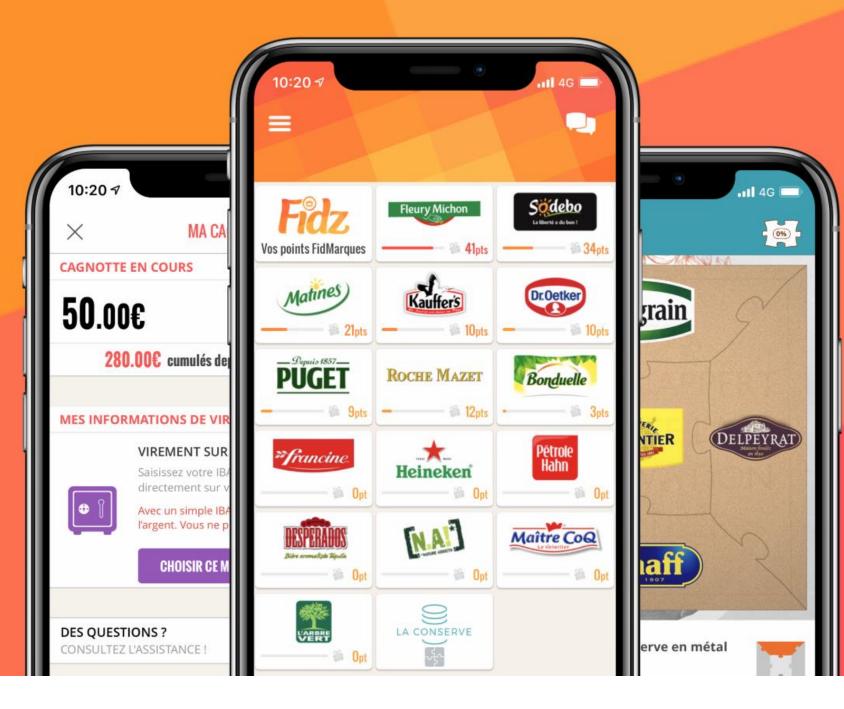
Fid Marques

L'appli gratuite qui récompense votre fidélité aux marques du quotidien !

É Télécharger sur iOS



Under the hood of receipt extraction

June 2019

SIFED 2019

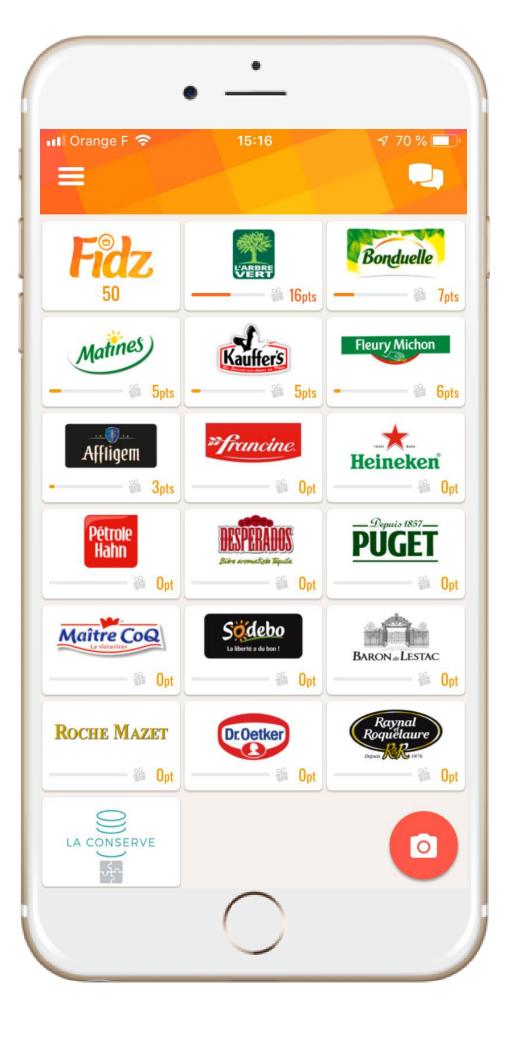
Télécharger sur Android



PurchEase







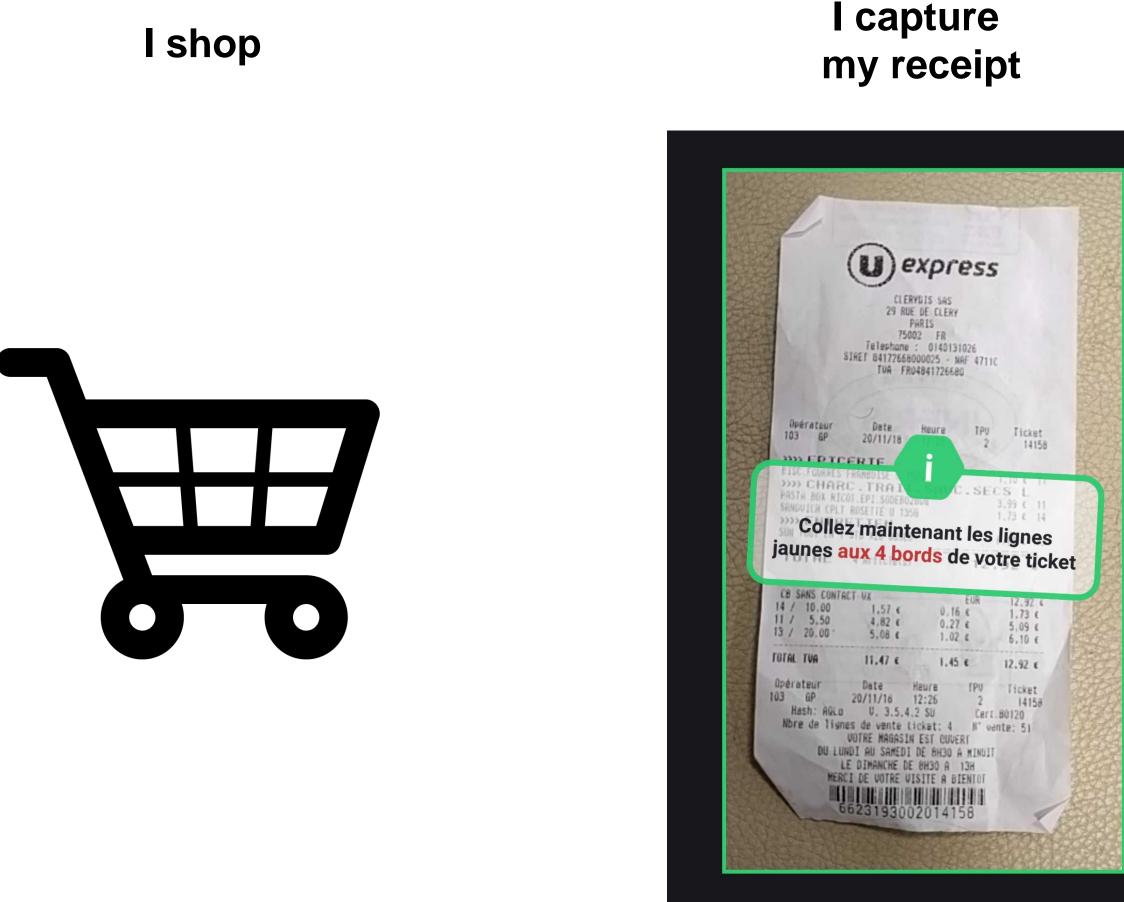
- 2- Receipt Extraction : a corner stone
- 3- Tech blog article vs real life
- 4- Summary of our stack
- 5- Focus on training : synthetic image database construction
- 6- Live Demo

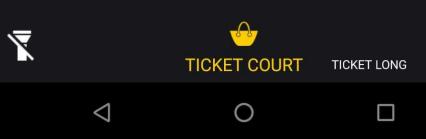
1- FidMarques : loyalty cards for every day groceries





FidMarques : loyalty cards for every day groceries



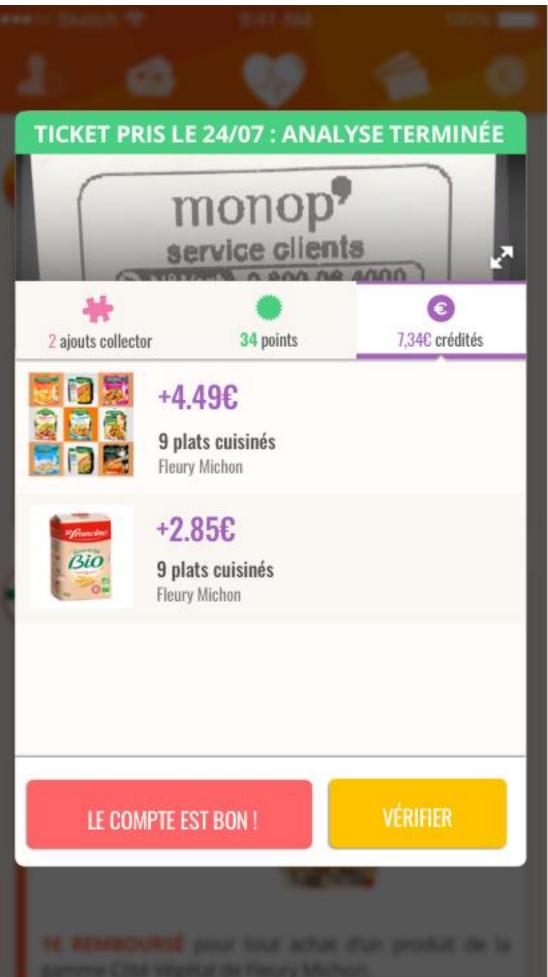


I earn Ioyalty points



FidMarques : loyalty cards for every day groceries

I burn points in exchange of refunds

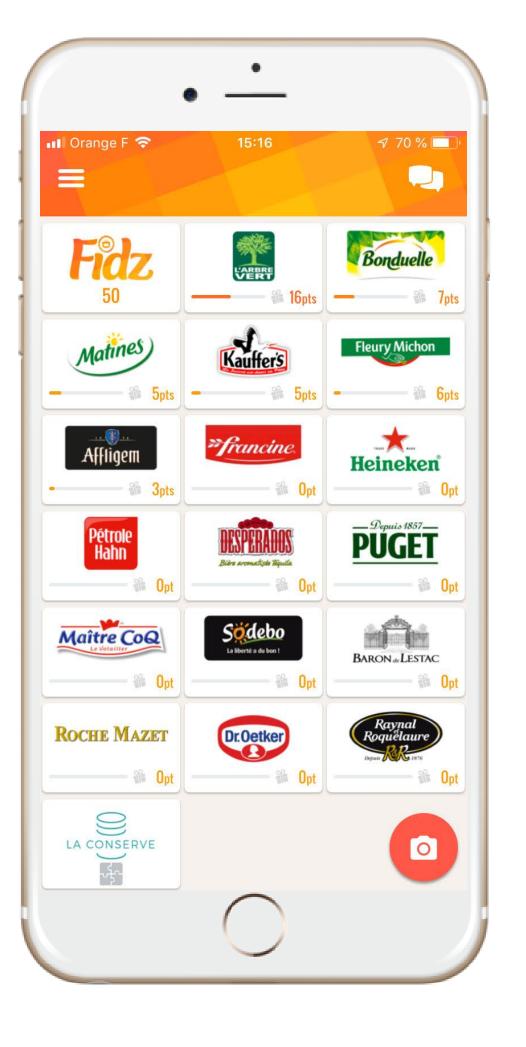




Steen C

Gamme Xtrem Box Validé le 08/01/2019





1- FidMarques : loyalty cards for every day groceries

2- Receipt Extraction : a corner stone

- 3- Tech blog article vs real life
- 4- Summary of our stack
- 5- Focus on training : synthetic image database construction
- 6- Live Demo



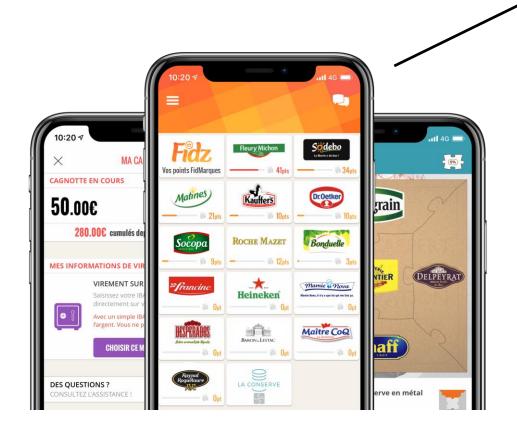


2- Receipt Extraction : a corner stone

Multi channel approach



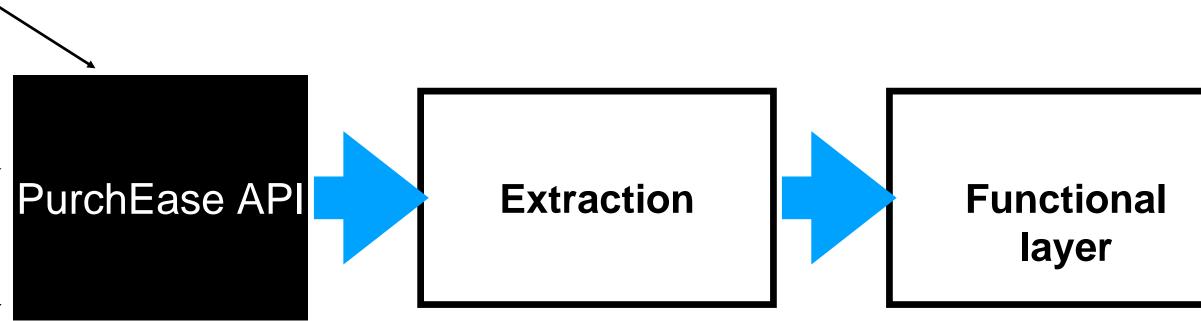
brand web site (no control on input)



(a)

Mobile apps (capture under control)

email (structured data in attachment or in body)



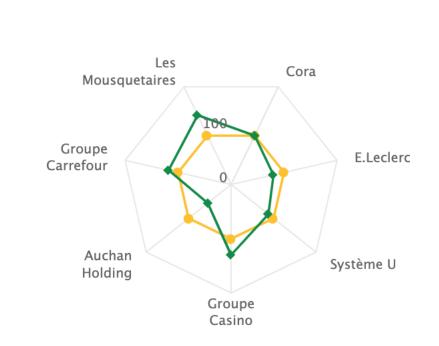


In a "B to B to C" model we need to be accurate for the end user, but also for the client

User constraints

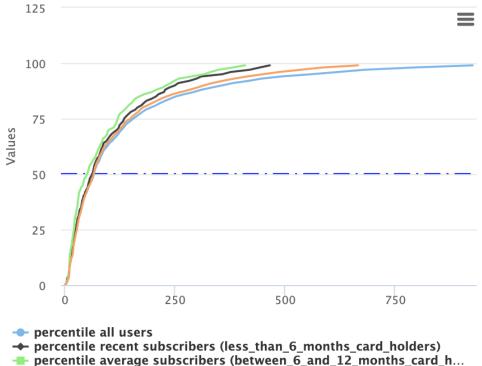
- if a receipt is merely readable it **must** be read
- rejections accepts no false positive

Exemples de données renvoyées a la marque



Tendance des ventes

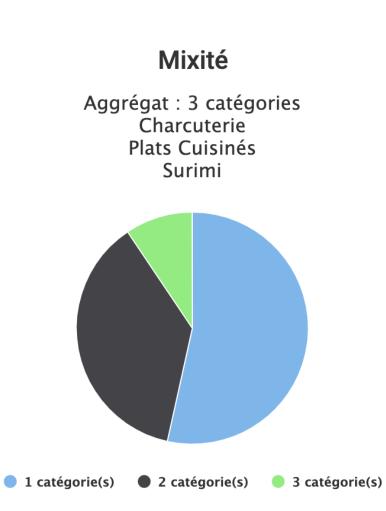
Percentile des dépenses annuelles

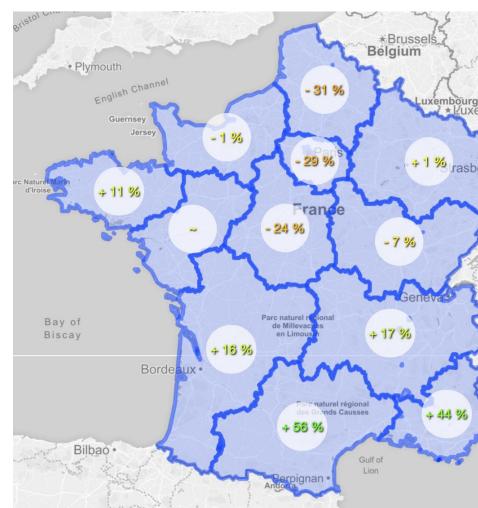


percentile old subscribers (more_than_12_months_card_holders)

Client constraints

- The whole receipt must be read
- metadata must be accurate
- point of sale discounts must be detected
- bought products must be identified beyond the brand
- fraudulent receipts must be dismissed

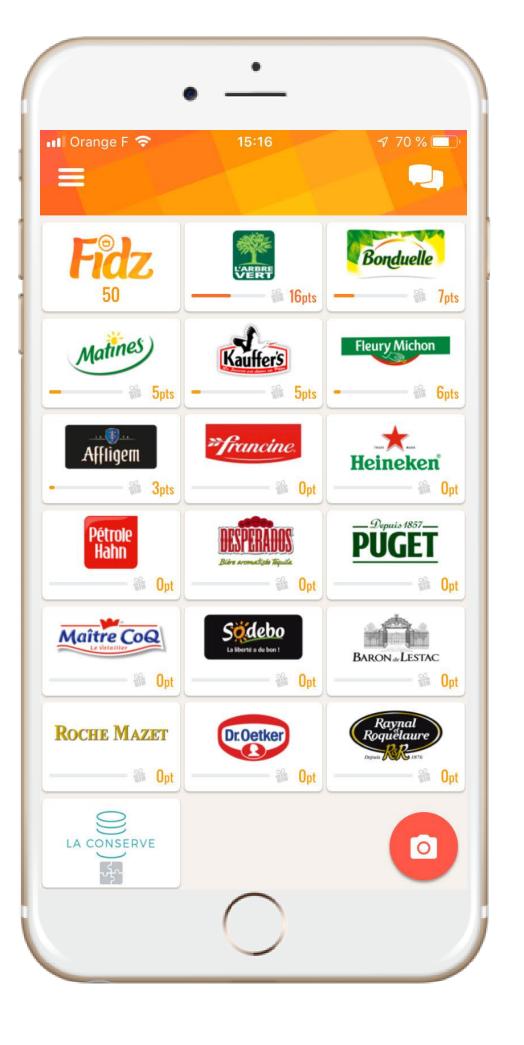












- 1- FidMarques : loyalty cards for every day groceries
- 2- Receipt Extraction : a corner stone
- **3- Tech blog article vs real life**
- 4- Summary of our stack
- 5- Focus on training : synthetic image database construction
- 6- Live Demo





3- Tech blog article vs real life (1)

Once every 6 months on hacker news you'll find an article with the following recipe :

•create a python script

- use a few openCV tricks for segmentation
- •get your hand on an OCR library (tesseract / google vision ...)
- apply a few regex
- •congrats, you have a receipt reading system !

3- Tech blog article vs real life (2)

However, life is not a bed of roses

The user can send whatever he wants:





Turn your head by 145 degree

Baby receipt

	VOTRE FACTURE SEPI	HORA	
	Vous venez de faire de magnifiques	acquisitions	
	Bonjour Odile,		
	Nous yous remembers gour votre commande N. 7882303779	manufactory and sectors for 1	
	La facture n' 1990/111985 en cole du 23 mai 2019 446 étable à vol-		
	Imprimer et de la conserver.		
	Commande web e* 7002308 Date de la vente ou de la prestation de ser	779. Kos du 28/06/2019	
	CLARIN 8 Baume Corps Super Hydratant		
	E Obiii Pricuriaire (TVS3000%/relaat)	57,00 EUR	
	Promotion Intel TTC	-14,25 EUR 42,76 EUR	
	Biggi Tru		
	Castanha do Brasil - Beure corp	orel	
	Gal 1 Prix untains (T)(3.20,00%/hc/user)	sr,de CUR	
Training of the second s	Promotion	-8,00 EUR	
1000 B2224 F B 2 7	Ibial TTC	18,00 EUR	
Olive Store 107	ORANADO		
	Caléndula - Beure corporei		
1000 10000000	Price unitable (TVS.00,00% Incluse) Promotion	3440 CUR -8.00 EUR	
	Todasi TTC	18,00 EUR	
	3 Actantilur(k) offert(k)*		
	Vien antidige		
	Erchallage Ladeaux	NON	
	Massaga Ladanu	NON	
	Fraix de livraixon (montent/TTC)	0.00 8.14	
	Government (managering)	ALA- DJR	
	74130.00%	YEAL BUR	
	Manuare de verre facture TTE	78,76 EUR	
	Décal de pulserent		
	Palanarran das dubli na 1918		
	Carla de stédit dés sourcesses entre-6017	1275 B/R	
	ADRESSE DE LIVRAJSON ADRE		
		and the second	
	1011005	One BIORS	

MASQUE SOS HTE EXIGENC LOTION TONI PINCE A EPI UAPORISATEI FLASHLINER TOTAL 6 BT	PA NOCIBE 34200 SELL 0467745884 V hats du 16/02/20.9 a 4				
СВ	01				
Code 5 20%00	a 0° ANNS BAUM FAUTH FELAL 38.50 €				
P	a)220. CLARIN CONTROL SER				
Points 1	= 58.50 € 1otal 97.00 €				
Solde F	a 8999/ 40% ∠nd pdt = -15.40 €				
Vous VE	Total 81.60 €				
UDUS VE	ancaire : 81.60 €				
Solde E					
	ady va Montant H.T. T.V.A. 201 % 68.00 € 13.60 €				
	votre espace fidelité				
	CATIA CLARUEL Cal fidélité no 21966582 Jour avez été accueillie par CELT				
	Votre ancien solde de points : 25 VOTRE NOUVEAU CUMUL POINTS : 106				
	Echange possible sur présentation de ce ticket, dans un emballage d origine en parfait état et dans un de 30 jours				
	ez votre flacon vide e. énéficiez de 20% de remise sur 1 achat d'un nouveau parfum. Voir conditions en magasin				
	MERCI DE VOTRE VISITE ET A BIENTU				

Oups! I put it in my pants

Today's digital age!





3- Tech blog article vs real life (2)



Oups! I do the math!

	100						
PHARMACIE LAFAY Mme Blanc M. Carpe 20 avenue Tolosane 31520 RAMONVILLO TEL : 05 61 73 20 7 FA180138245 COM Produit	E ST A le I (PTOIR4 Qté A	: 25/06/3 Op : 28 1 régler To	2018 3:10				
Vente du 25/06/18 1 pour Sophie SPORN BCOMBIO BIEN-E	80138245 J	5					
VEG GEL D 200ML	2	5,19	10,38				
Remise ligne,%	50,00		-5,19				
ONCTOSE H.C.CR TB30G	1	4,10	4,10				
LRP MINI EAU							
MICELL EFFACLAR	4	4,90	19,60				
100ML QUIES BOULE CIRE							
BT24	1	3,93	3,93				
EFFACLAR MAT SO	Ν						
HYDRATANT	1	9,75	9,75				
SEBOREGUL CREALINE H2O EAU							
NETT 2X500ML	1	16,42					
Sous total remisé en l	EUR		58,99				
Net à payer en	EUR	UR = 6	58,99 ,55957 F				
Net	à payer	enF=	386,95 F				
Dont TI/A .			0,04				
asing porte Carte Bancaire en cur de, de							
New Justiclos '							
Montant des Honor, de Disp.							
Nombre de points	de Fidél	ité : 12	2				

Difficult to interpret, even for human!

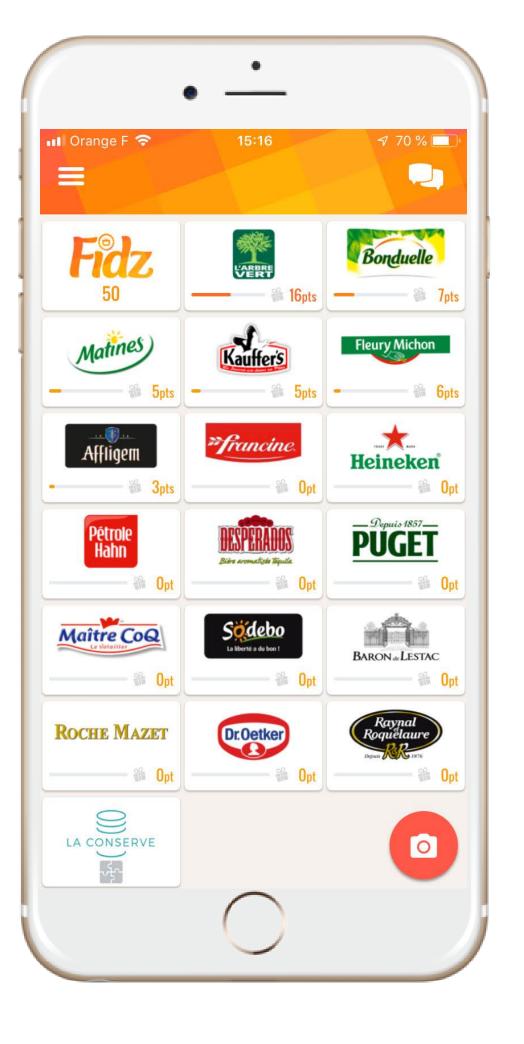
3- Tech blog article vs real life (3)

In summary:

- Unknown camera devices
- No information about the uploaded photo, such as :
 - Image resolution
 - Receipt length
 - Receipt position, orientation
 - Number of receipts
 - Background

- Many distortions can occur during image acquisition :
 - Out of focus
 - Un uniform light
 - High-ISO noise
- Wrapped / curved / folded paper problem
- Degraded paper, vanishing text problems
- Etc.



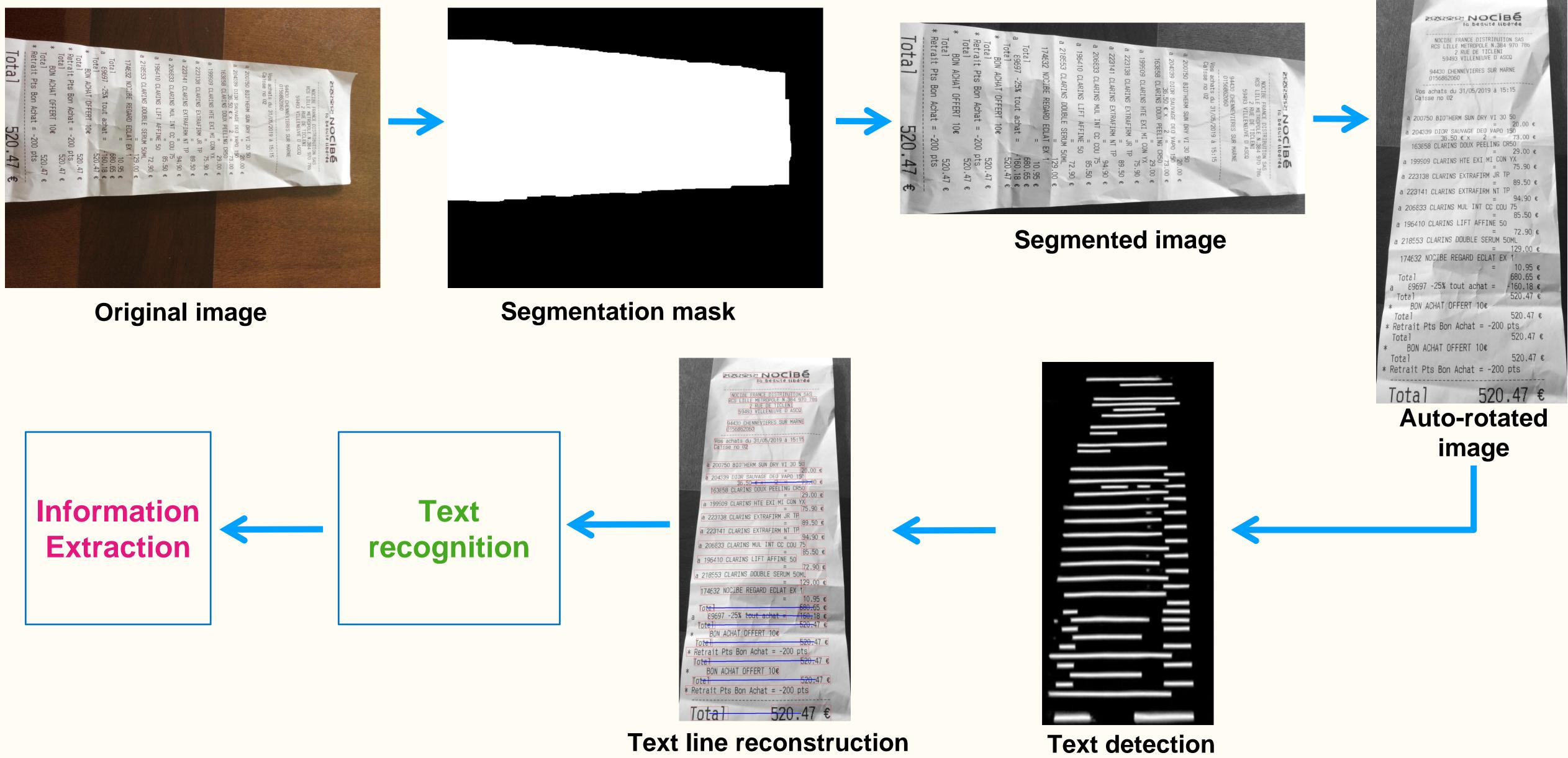


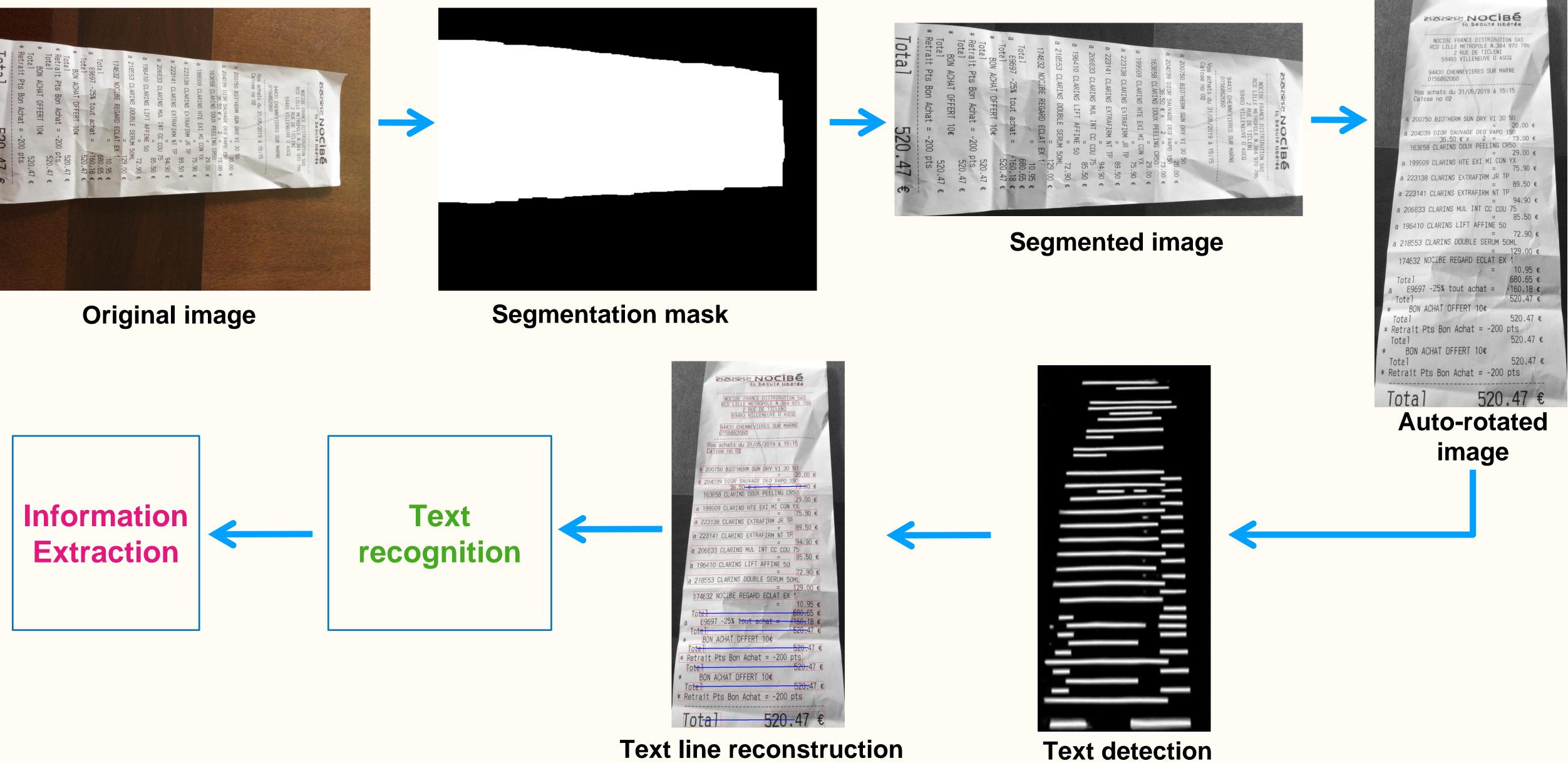
- 1- FidMarques : loyalty cards for every day groceries
- 2- Receipt Extraction : a corner stone
- 3- Tech blog article vs real life
- **4- Summary of our stack**
- 5- Focus on training : synthetic image database construction
- 6- Live Demo





4 Summary of our stack





4.1 Receipt detection and segmentation

1. Previous methods:

- Contour detection + Hough transform
- Split & Merge
- K-means
- Watershed

Pros:

- Simple to implement
- Usually fast

Cons:

- In reality : usually fail when receipt is not always homogenous white rectangle region in the image
- Due to noise / deformations \rightarrow poor results

2. Current method : based on convolutional network with dilated convolutions

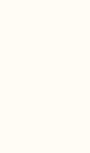
Pros:

- Features are discovered and extracted through training.
- Complex calculations \rightarrow Handle well the high variability in the input data
- Better result

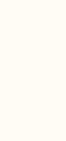
Cons:

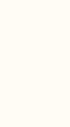
- Requires labelled (annotated) data for training process
- Usually slow













4.2 Skew angle detection and correction

1. Previous methods:

- Hough transform lacksquare
- Radon transform
- Texture analysis ullet
- Run-lengths of binary image objects ullet

Pros:

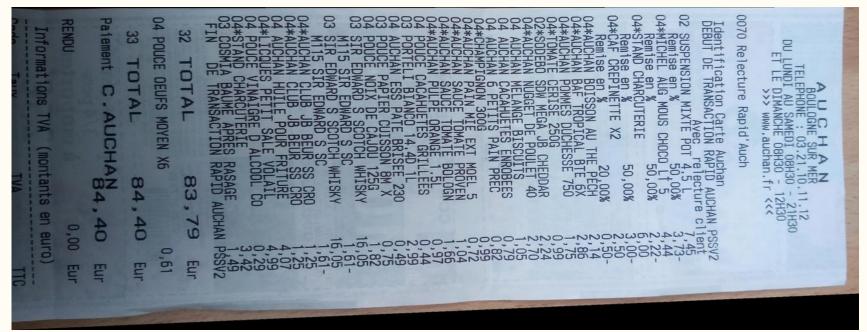
- Simple to implement
- Usually fast -

Cons:

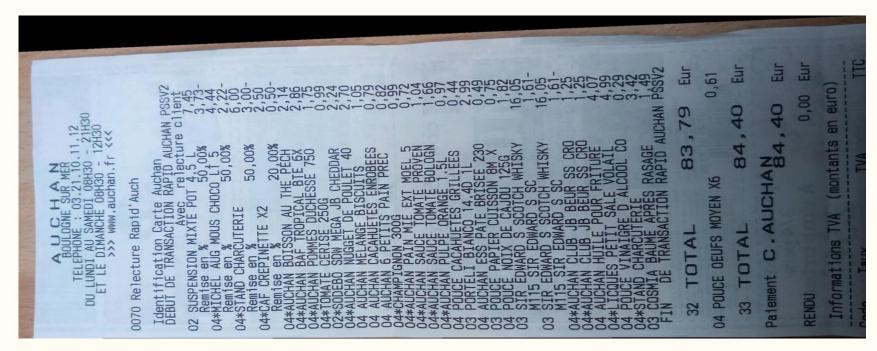
- Usually fail in the case presented in the figure
- **2. Current method :** based on convolutional neural network **Pros**:
- Complex calculations \rightarrow Handle well the high variability in the input data
- Better result
- Can handle very well problems presented in the figure

Cons:

- Requires annotated training data
- Usually slower



Which angle?



Which angle?

4.3 Text detection

1. Previous methods:

Morphological operations

Pros:

- Simple to implement
- Usually fast

Cons:

- Noises \rightarrow poor results
- Folded / wrapped / curved paper \rightarrow texts are not always straight lines \rightarrow poor results

2. Current method : based on fully convolutional neural network

Pros:

- Noisy input images : better result
- Handle very well folded / curved paper problem Cons:
- Requires labelled (annotated) data for training process





Result of current method

4.4. Text line reconstruction

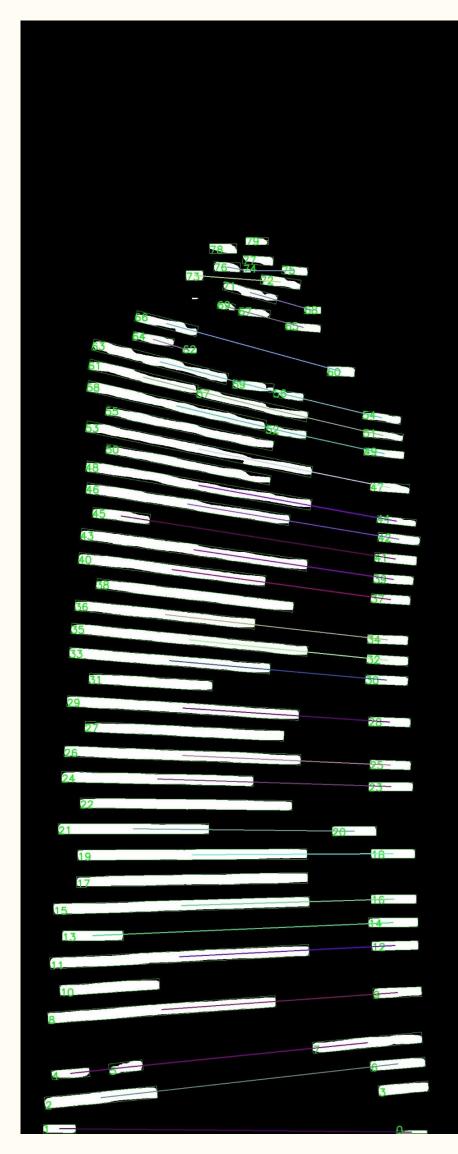
Objective:

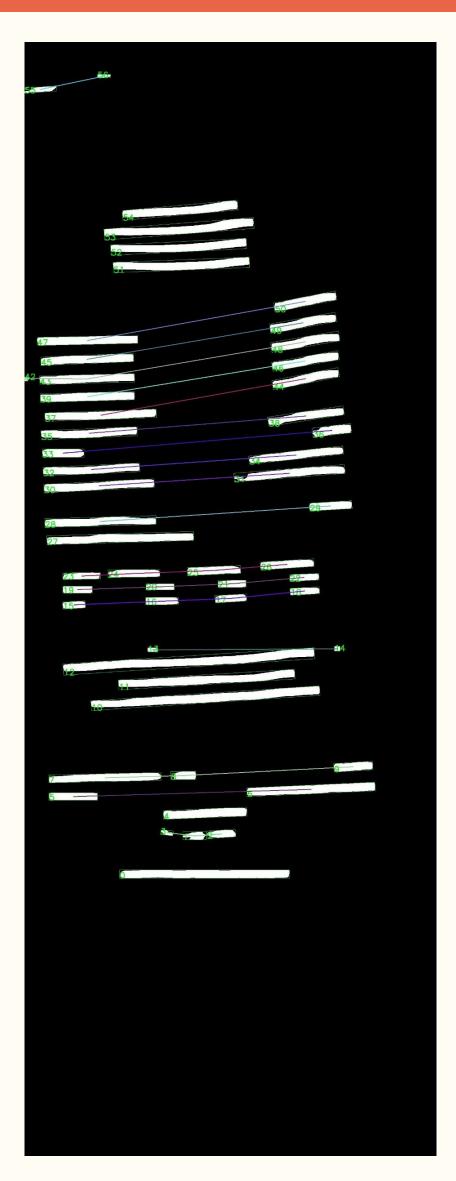
Group detected text objects (textons) into text lines

Methods:

- Greedy algorithm
- Genetic algorithm







Some results of text line reconstruction

4.5. Text line recognition & Information Extraction

Text line recognition:

Method: CNN + LSTM + CTC



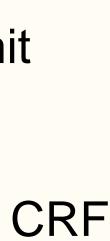
Text line recognition result

Low level Information Extraction :

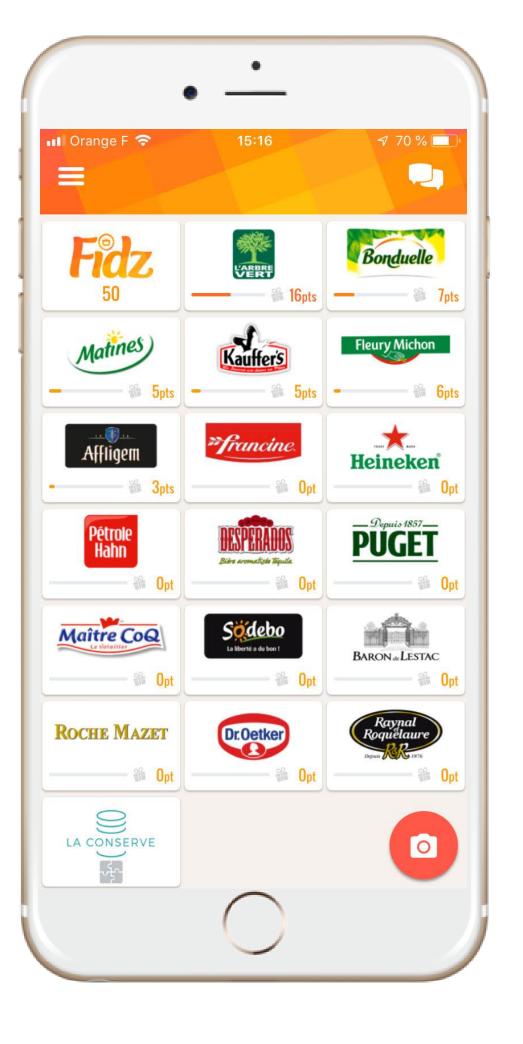
- Purchase information: product name, product's unit price, quantity, purchase date, ...
- Store's name, address, telephone number, ...
- Method: Named Entity Recognition using LSTM + CRF

SHNP HYPER U						
C C DU BOIS CANY						
GRAND QUEVILLY						
ADDR : 76120 FR						
PHONE : 0235183434						
siret : 353185952000	69					
TVA FR96353185952						
Operateur Date Heure TPV Ticket						
DATE :	2019-06-01	18:33				
>>>> PAPIER TOILETTE						
PAPIER TOIL MOLT LOTUS 18RLXFF	1x6.99	6.99				
>>>> EAUX						
EAU SOURCE CRISTALINE 6X1 5L	2x0.99	1.98				

Result from low level information extraction







- 1- FidMarques : loyalty cards for every day groceries
- 2- Receipt Extraction : a corner stone
- 3- Tech blog article vs real life
- 4- Summary of our stack
- 5- Focus on training : synthetic image database construction
- 6- Live Demo





5. Synthetic image generation

In section 4:

- Neural networks is powerful •
- But it requires high amount of annotated data \bullet

To obtain annotated data:

- Hire human workers
- However:
 - Require many \$\$\$\$\$\$
 - Lack of qualified human resource

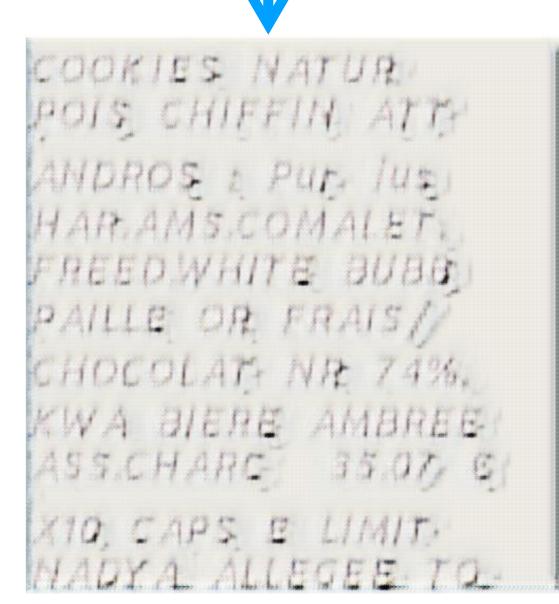
Our solution:

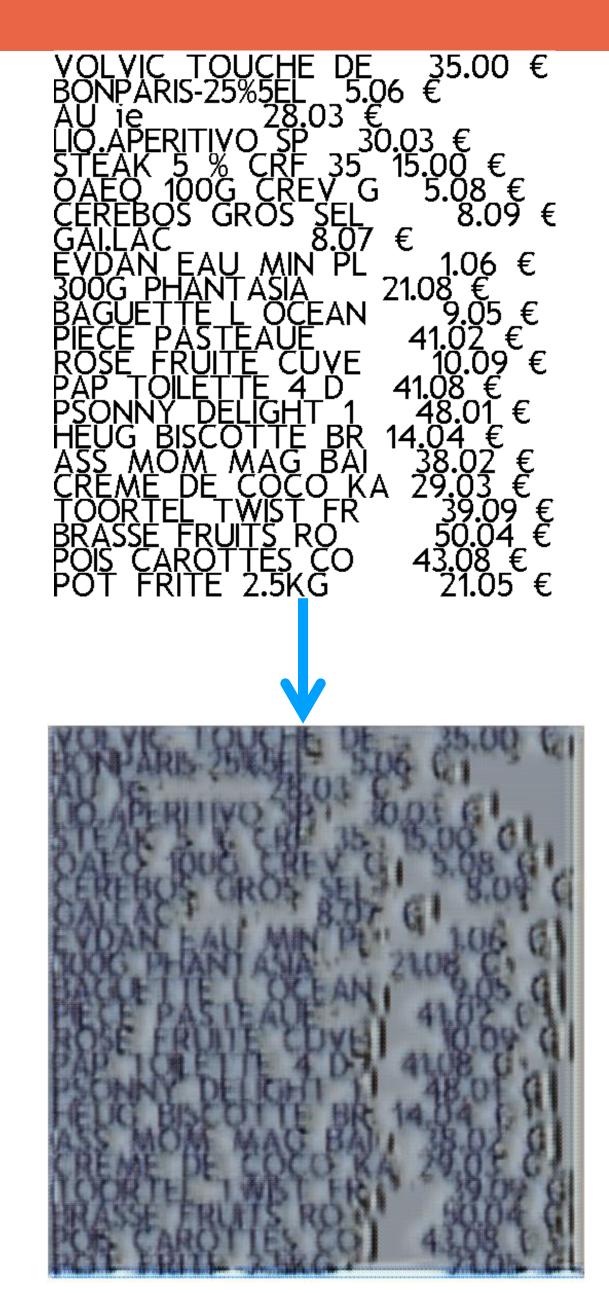
- Synthetic image generation using GAN
- Based on image domain translation -
- distortions

Idea: Distortions as style. Translate Binary images with no distortion \rightarrow Realistic images with

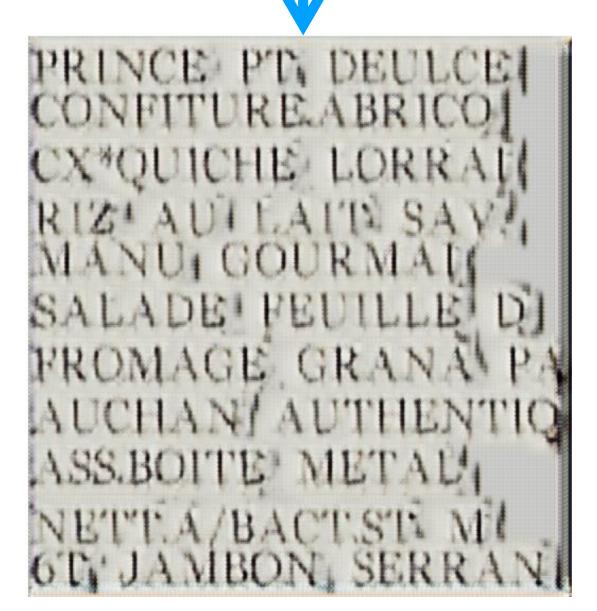
5. Synthetic image generation - Examples

COOKIES NATUR POIS CHIFFIN ATT ANDROS : Pur ius HAR.AMS.COMALET. FREED.WHITE BUBB PAILLE OR FRAIS/ CHOCOLAT NR 74%. KWA BIERE AMBREE ASS.CHARC 35.07 € X10 CAPS E LIMIT NADYA ALLEGEE TO



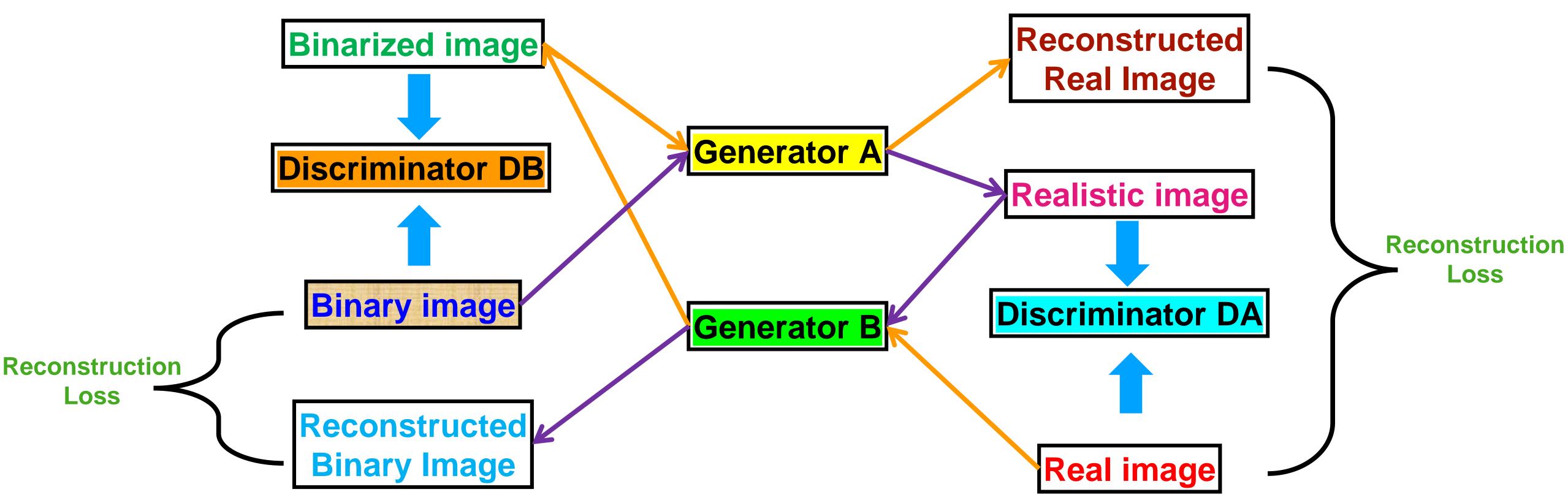


PRINCE PT DEULCE CONFITURE.ABRICO CX*QUICHE LORRAI RIZ AU LAIT SAV. MANU GOURMAI SALADE FEUILLE D FROMAGE GRANA PA AUCHAN AUTHENTIQ ASS.BOITE METAL NETT.A/BACT.ST M 6T JAMBON SERRAN



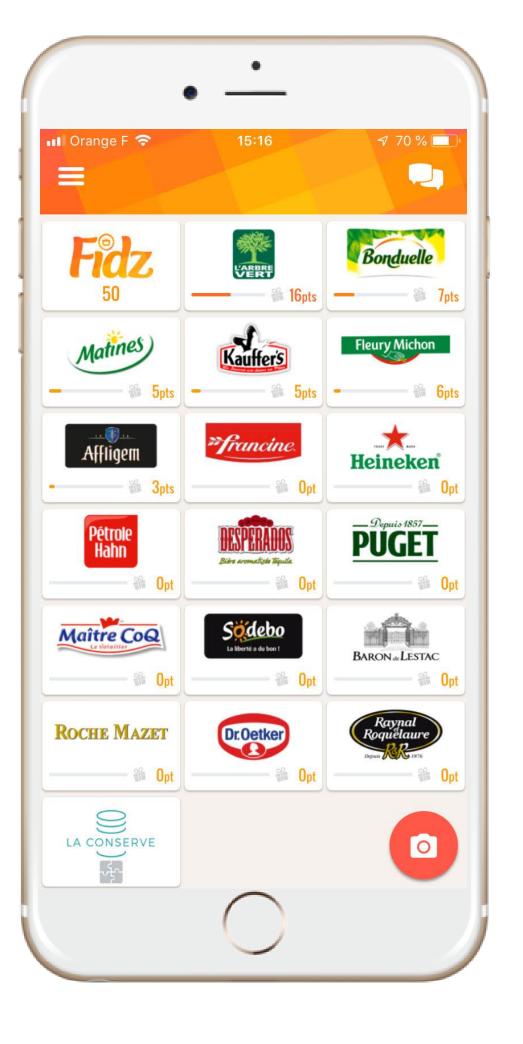
5. Synthetic image generation

DualGAN:



Loss





- 1- FidMarques : loyalty cards for every day groceries
- 2- Receipt Extraction : a corner stone
- 3- Tech blog article vs real life
- 4- Summary of our stack
- 5- Focus on training : synthetic image database construction
- 6- Live Demo



