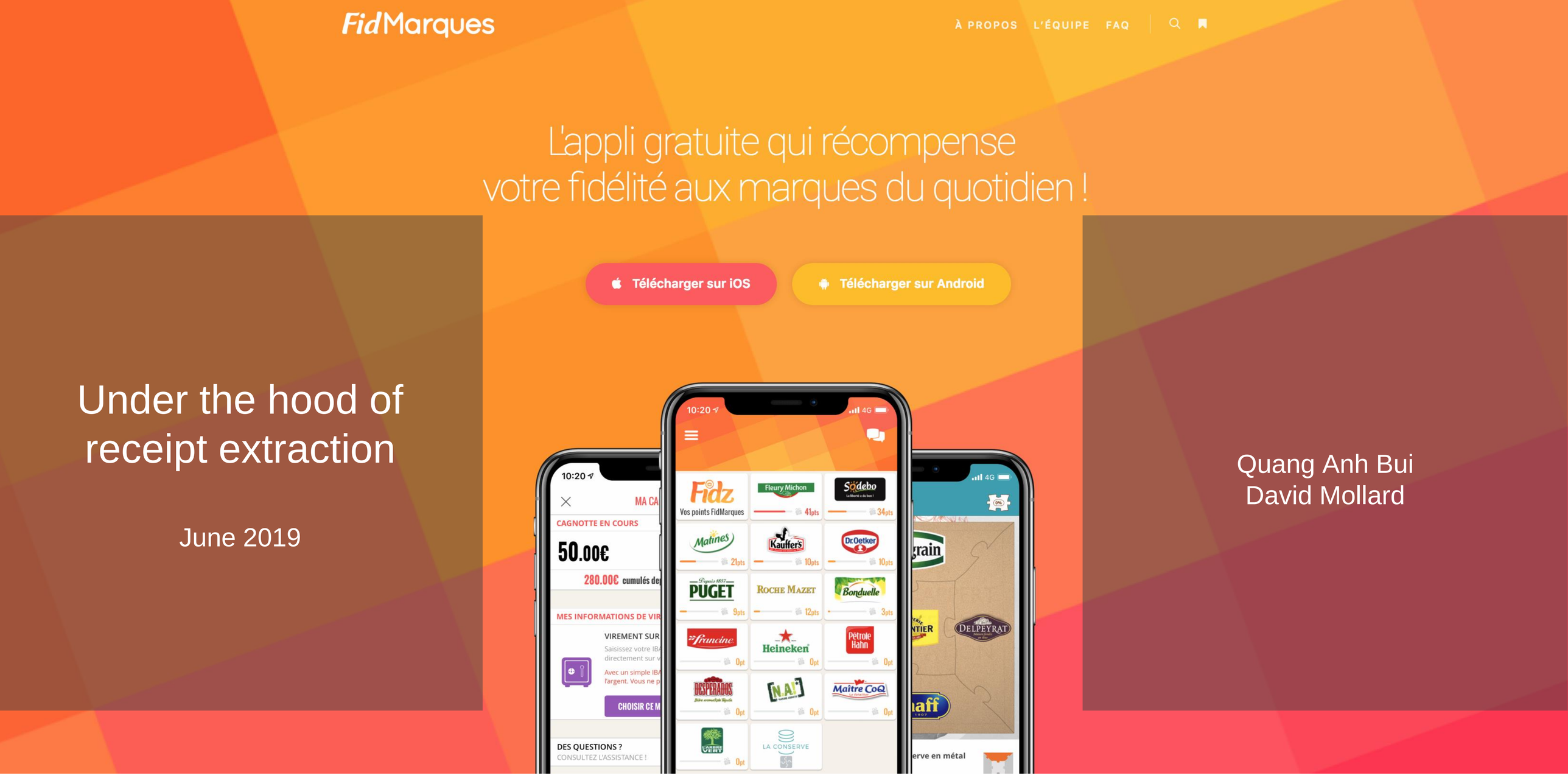
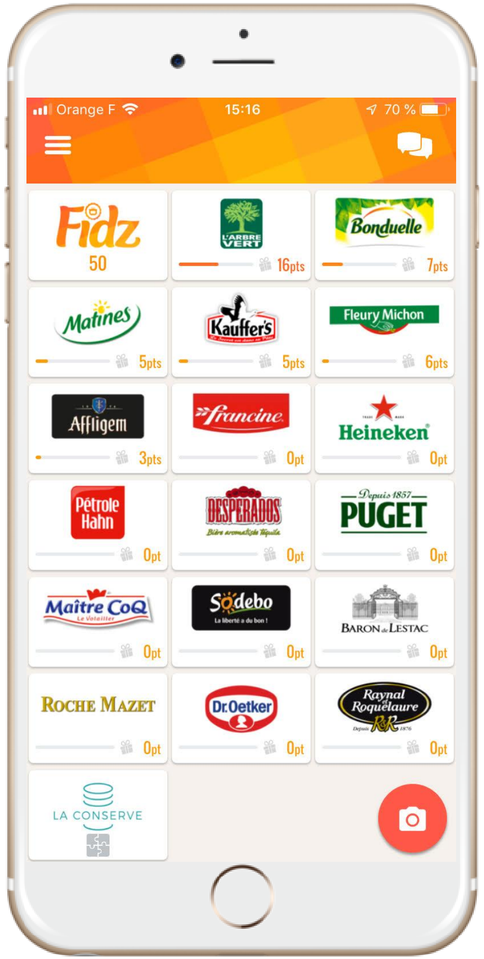


Under the hood of receipt extraction

June 2019

Quang Anh Bui
David Mollard

Agenda



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

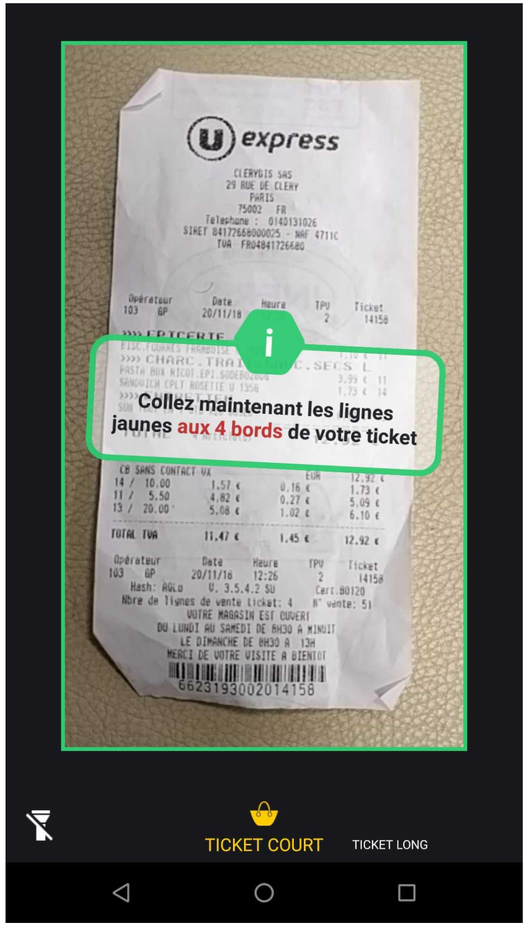
6- Live Demo

FidMarques

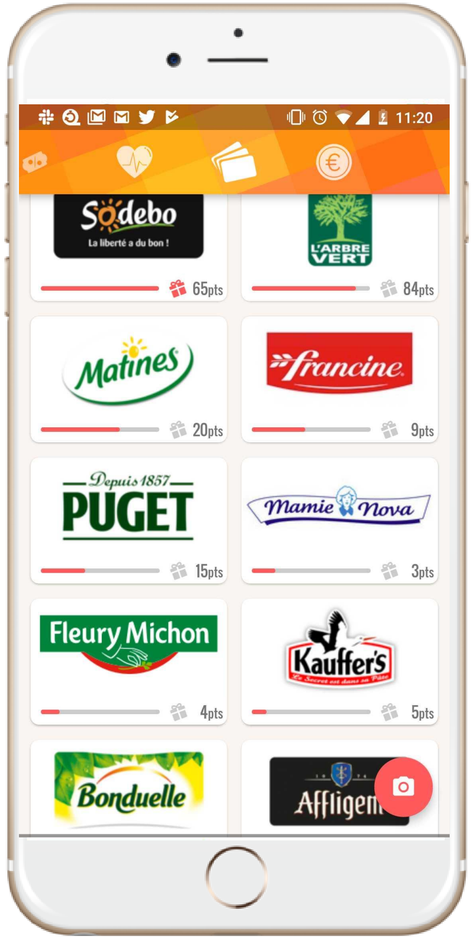
FidMarques : loyalty cards for every day groceries

I shop

I capture my receipt

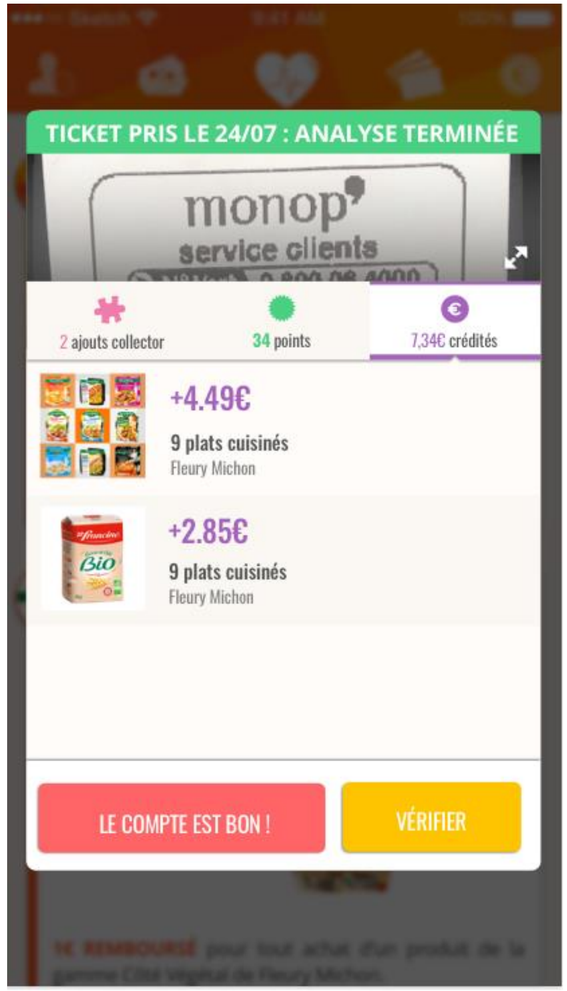


I earn loyalty points

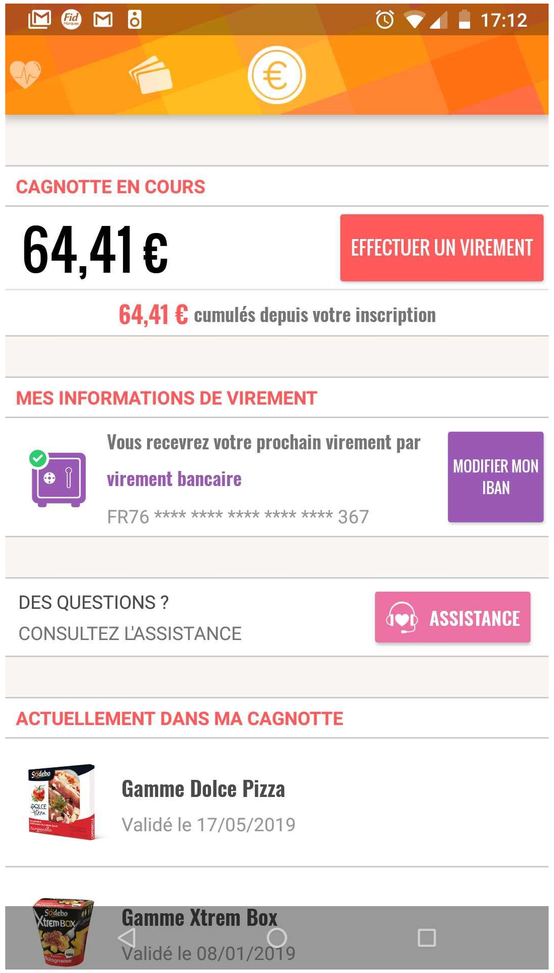


FidMarques : loyalty cards for every day groceries

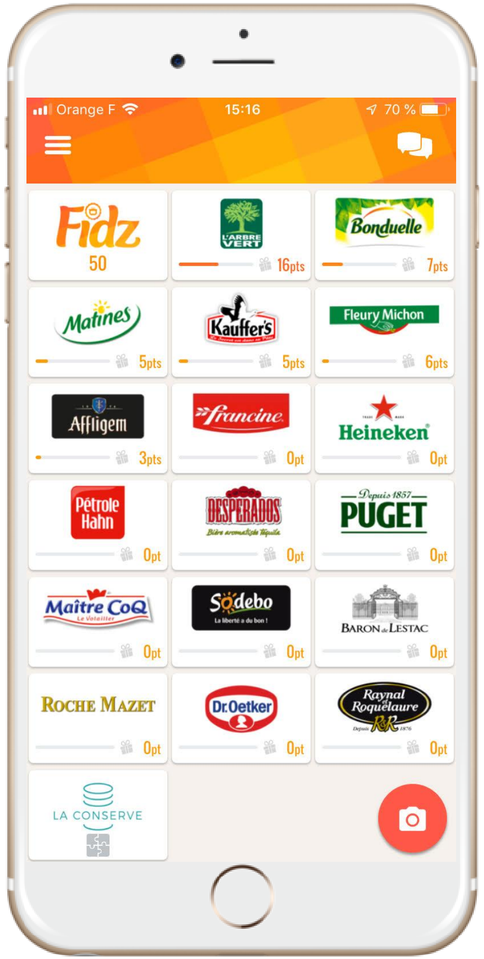
I burn points in exchange of refunds



I earn money 😀



Agenda



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

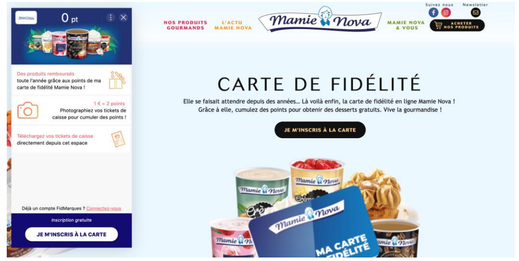
6- Live Demo

FidMarques

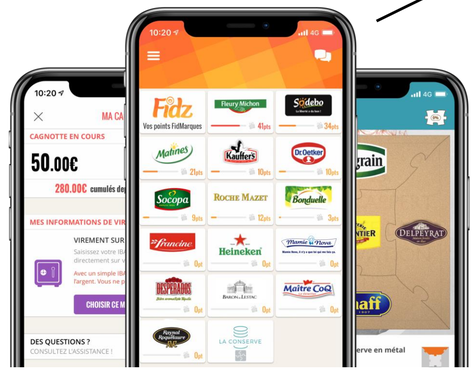
2- Receipt Extraction : a corner stone

Multi channel approach

email
(structured data
in attachment or in body)



brand web site
(no control on input)



Mobile apps (capture under control)

PurchEase API

Extraction

Functional layer

2- Receipt Extraction : a corner stone

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

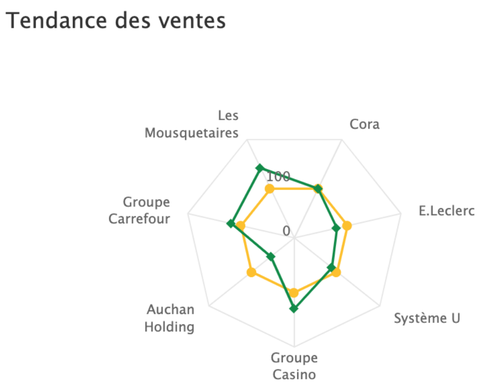
User constraints

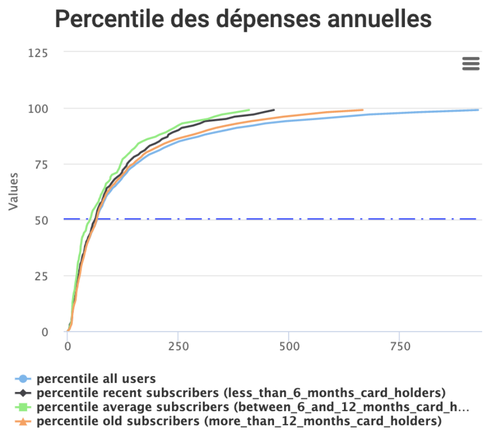
- if a receipt is merely readable it **must** be read
- rejections accepts no false positive

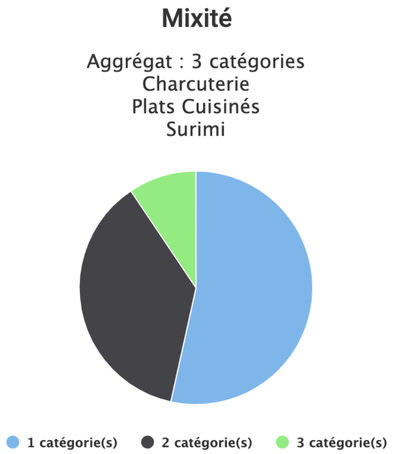
Client constraints

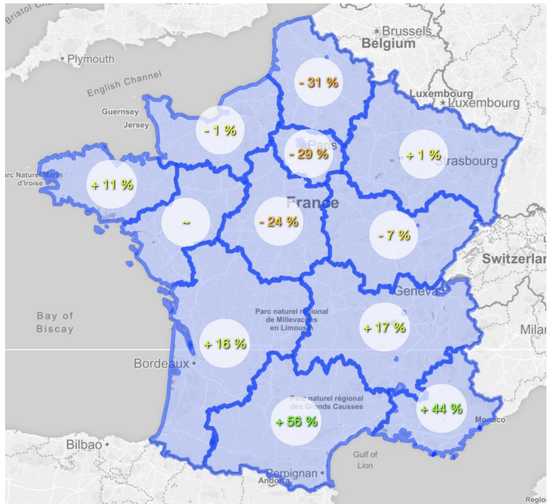
- The whole receipt must be read
- metadata must be accurate
- point of sale discounts must be detected
- bought products must be identified beyond the brand
- fraudulent receipts must be dismissed

Exemples de données renvoyées a la marque

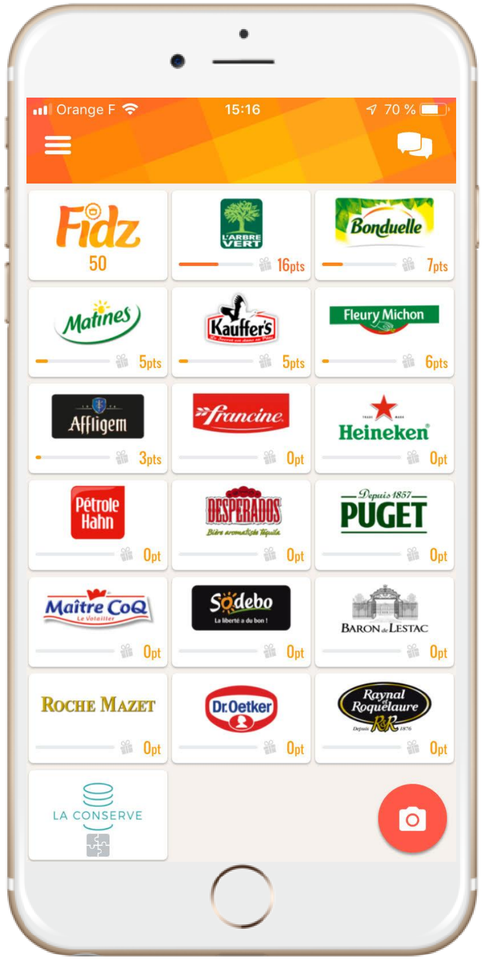








Agenda



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

FidMarques

3- Tech blog article vs real life (1)

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

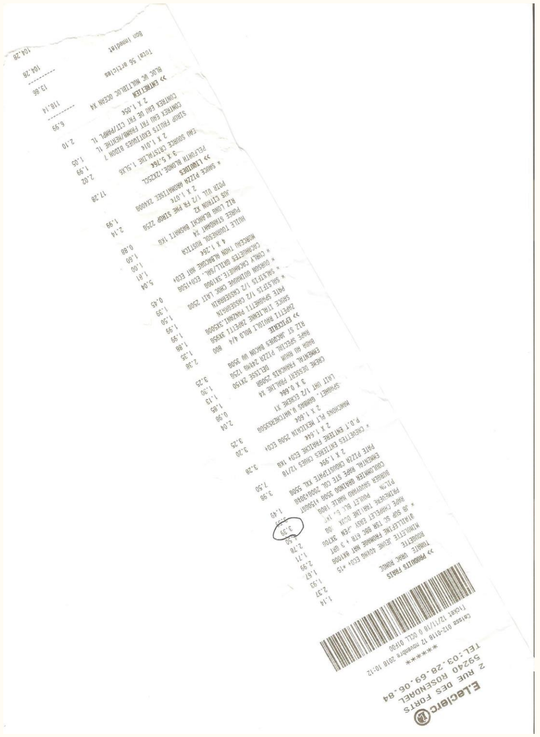
- **create a python script**
- **use a few openCV tricks for segmentation**
- **get your hand on an OCR library (tesseract / google vision …)**
- **apply a few regex**
- **congrats, you have a receipt reading system !**

3- Tech blog article vs real life (2)

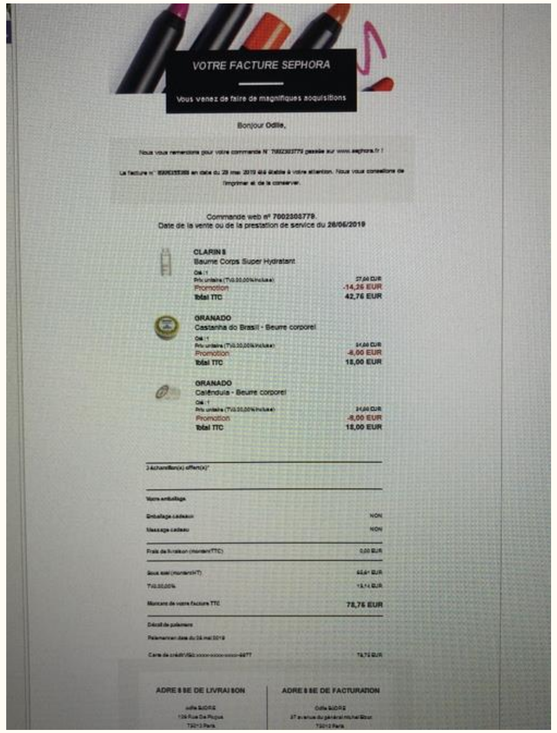
However, life is not a bed of roses

The user can send whatever he wants:

Baby receipt



Turn your head by 145 degree

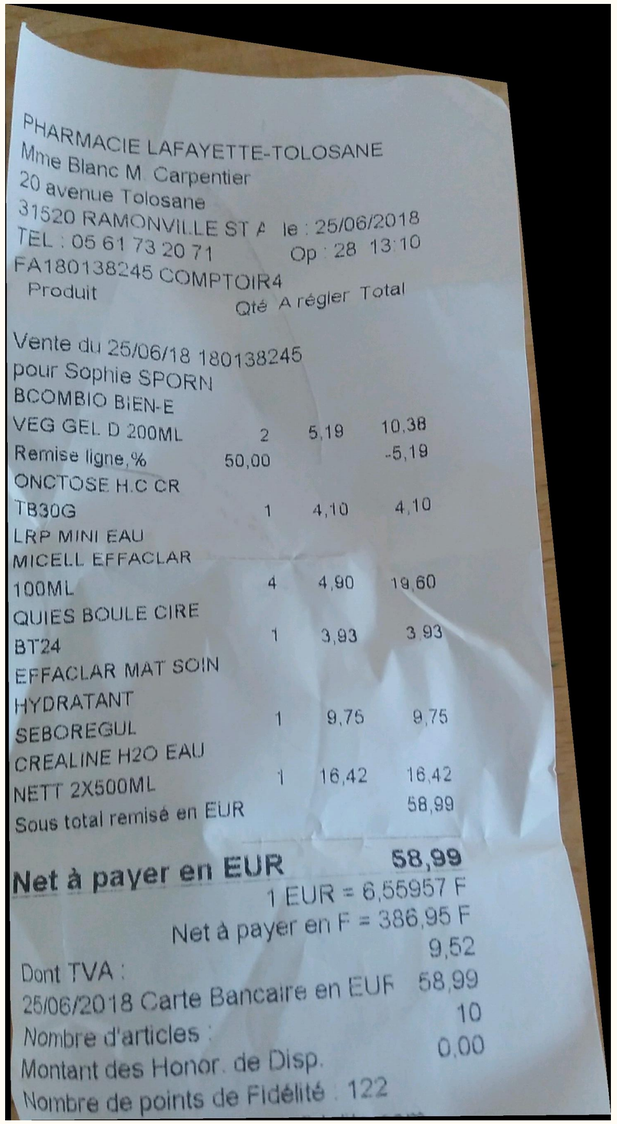


Today's digital age!

Oups! I put it in my pants

3- Tech blog article vs real life (2)

Oups! I do the math!



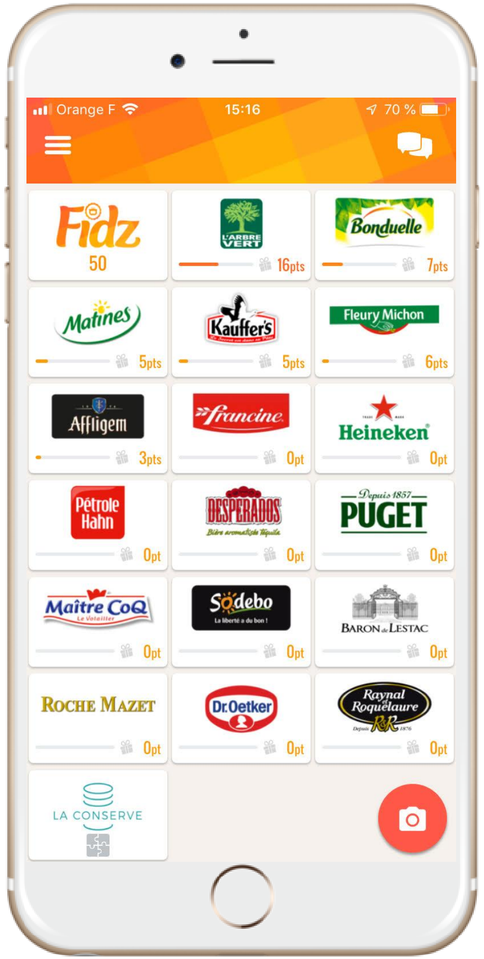
Difficult to interpret, even for human!

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.

Agenda



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

FidMarques

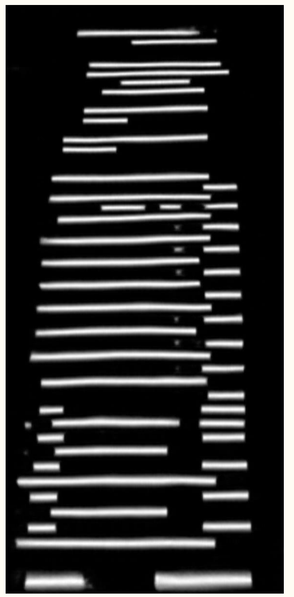
4 Summary of our stack

Original image

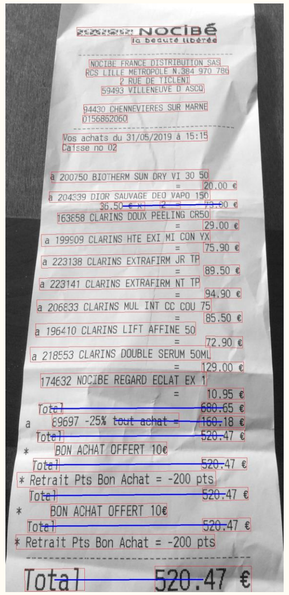
Segmentation mask

Segmented image

Auto-rotated image



Text detection



Text line reconstruction

Text recognition

Information Extraction

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 → poor results

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

Pros:

- Features are discovered and extracted through training.
- Complex calculations → 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:

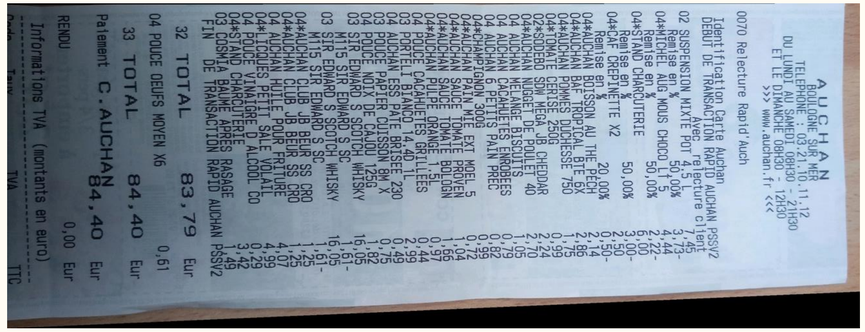
- Hough transform
- Radon transform
- Texture analysis
- Run-lengths of binary image objects

Pros:

- Simple to implement
- Usually fast

Cons:

- Usually fail in the case presented in the figure



Which angle?

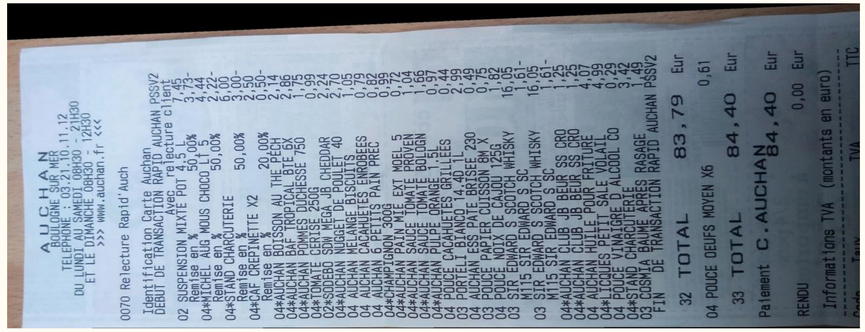
2. Current method : based on convolutional neural network

Pros:

- Complex calculations → 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?

4.3 Text detection

1. Previous methods:

- Morphological operations

Pros:

- Simple to implement
- Usually fast

Cons:

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

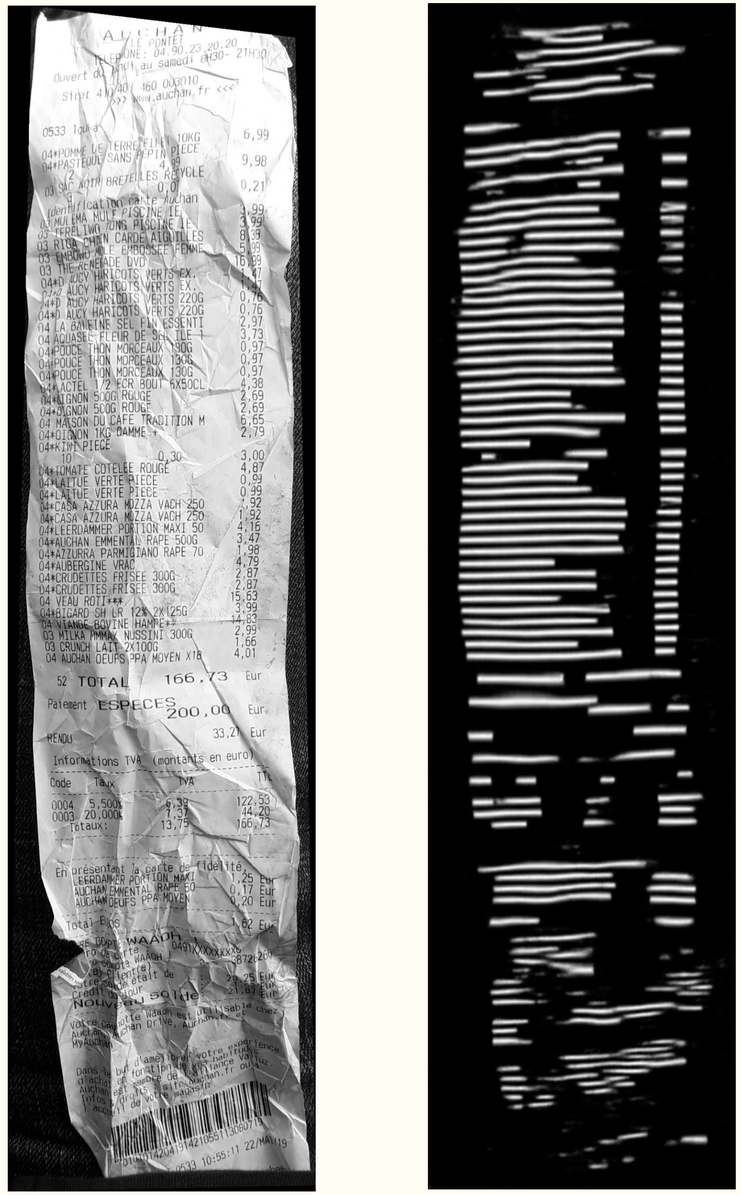
2. Current method : based on fully convolutional neural network

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

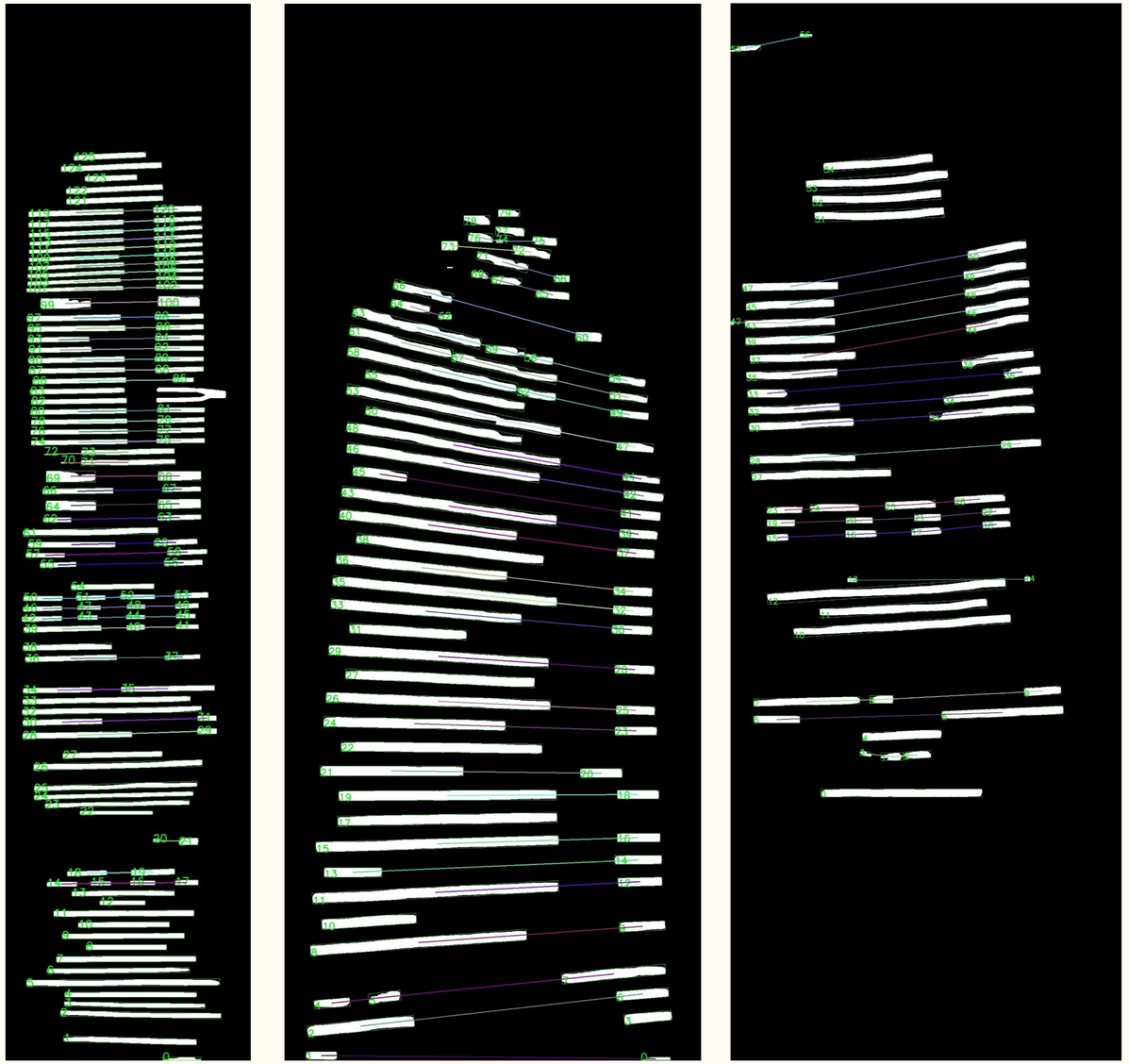
4.4. Text line reconstruction

Objective:

- Group detected text objects (textons) into text lines

Methods:

- Greedy algorithm
- Genetic algorithm

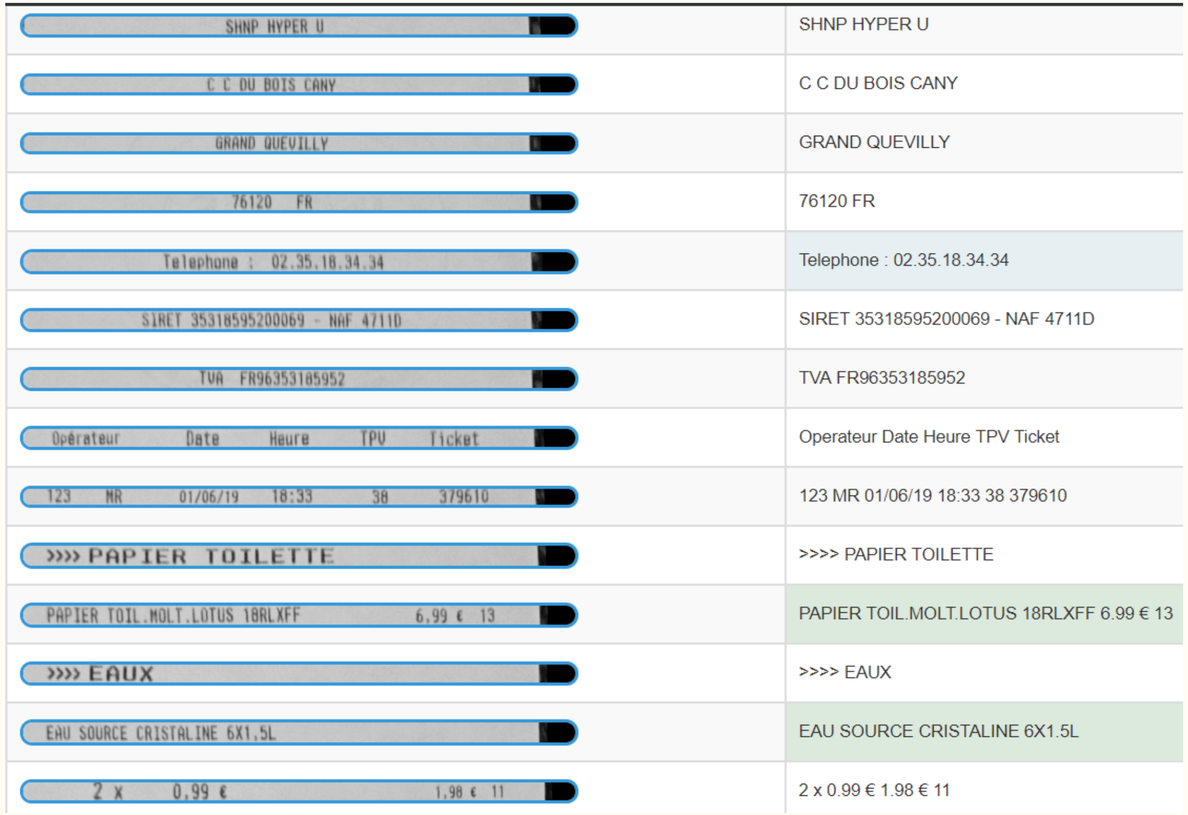


Some results of text line reconstruction

4.5. Text line recognition & Information Extraction

Text line recognition:

- Method: CNN + LSTM + CTC

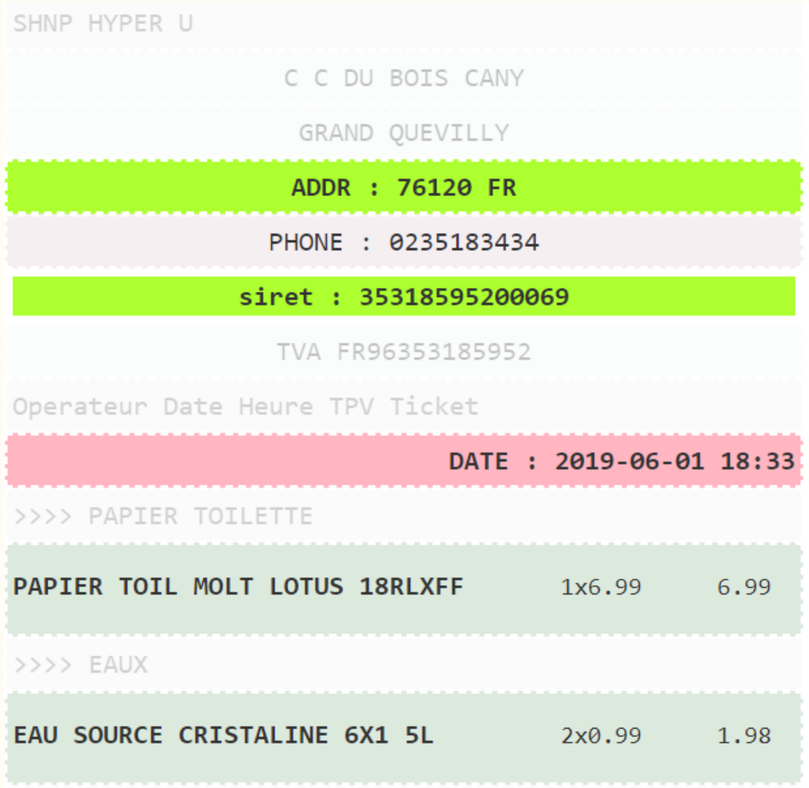


SHNP HYPER U	SHNP HYPER U
C C DU BOIS CANY	C C DU BOIS CANY
GRAND QUEVILLY	GRAND QUEVILLY
76120 FR	76120 FR
Telephone : 02.35.18.34.34	Telephone : 02.35.18.34.34
SIRET 35318595200069 - NAF 4711D	SIRET 35318595200069 - NAF 4711D
TVA FR96353185952	TVA FR96353185952
Opérateur Date Heure TPV Ticket	Operateur Date Heure TPV Ticket
123 MR 01/06/19 18:33 38 379610	123 MR 01/06/19 18:33 38 379610
>>>> PAPIER TOILETTE	>>>> PAPIER TOILETTE
PAPIER TOIL.MOLT.LOTUS 18RLXFF 6,99 € 13	PAPIER TOIL.MOLT.LOTUS 18RLXFF 6.99 € 13
>>>> EAUX	>>>> EAUX
EAU SOURCE CRISTALINE 6X1.5L	EAU SOURCE CRISTALINE 6X1.5L
2 x 0,99 € 1,98 € 11	2 x 0.99 € 1.98 € 11

Text line recognition result

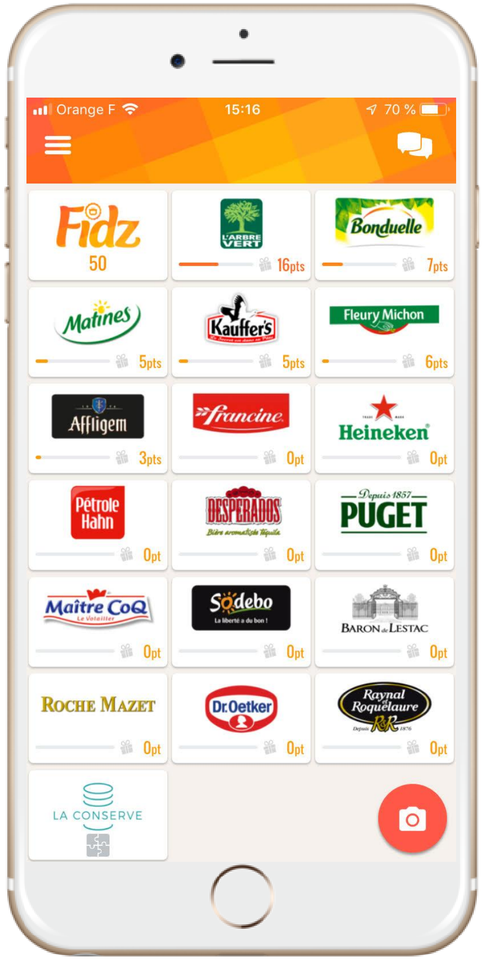
Low level Information Extraction :

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



Result from low level information extraction

Agenda



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

FidMarques

5. Synthetic image generation

In section 4:

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

To obtain annotated data:

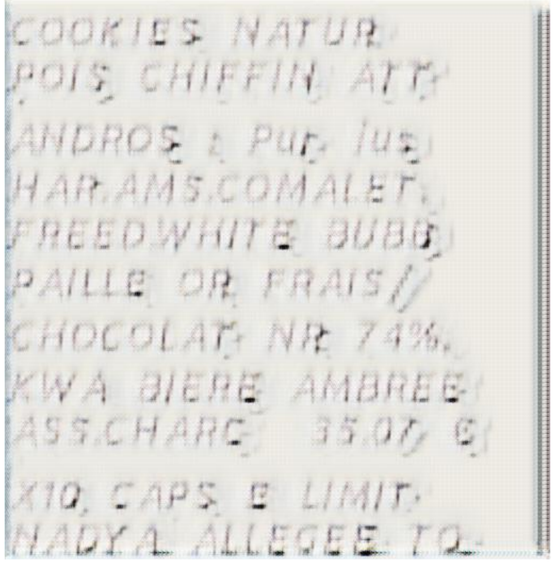
- Hire human workers
- However:
  - Require many $$$$$$$
  - Lack of qualified human resource

Our solution:

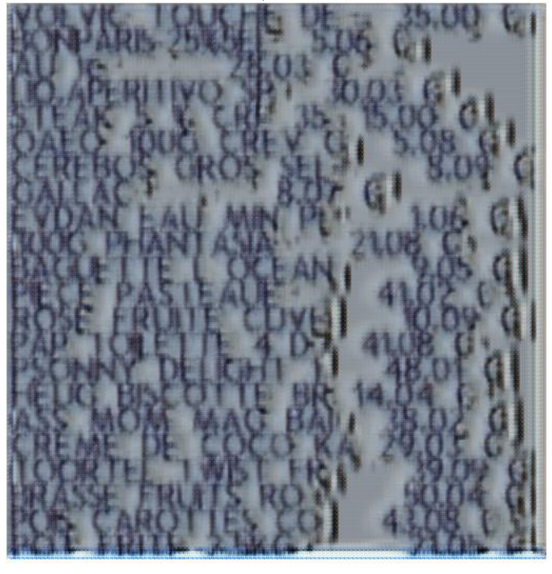
- Synthetic image generation using GAN
- Based on image domain translation
- Idea: Distortions as style. Translate Binary images with no distortion → Realistic images with distortions

5. Synthetic image generation - Examples

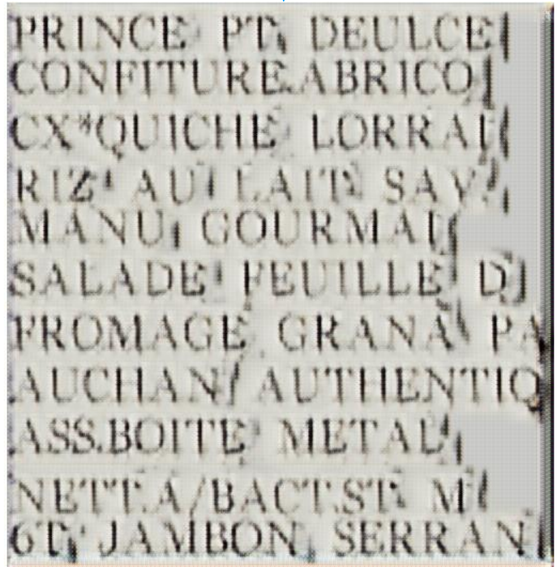
COOKIES NATUR
POIS CHIFFIN ATT
ANDROS : Pur ius
HAR.AMS.COMALET.
FREED.WHITE BUBB
PAILLE OR FRAIS/
CHOCOLAT NR 74%.
KWA BIERE AMBREE
ASS.CHARC 35.07 €
X10 CAPS E LIMIT
NADYA ALLEGEE TO



VOLVIC TOUCHE DE 35.00 €
BONPARIS-25%5EL 5.06 €
AU ie 28.03 €
LIQ.APERITIVO SP 30.03 €
STEAK 5 % CRF 35 15.00 €
OAEO 100G CREV G 5.08 €
CEREBOS GROS SEL 8.09 €
GAI.LAC 8.07 €
EVIAN EAU MIN PL 1.06 €
300G PHANTASIA 21.08 €
BAGUETTE L OCEAN 9.05 €
PIECE PASTEAUE 41.02 €
ROSE FRUITE CUVE 10.09 €
PAP TOILETTE 4 D 41.08 €
PSONNY DELIGHT 1 48.01 €
HEUG BISCOTTE BR 14.04 €
ASS MOM MAG BAI 38.02 €
CREME DE COCO KA 29.03 €
TOORTEL TWIST FR 39.09 €
BRASSE FRUITS RO 50.04 €
POIS CAROTTES CO 43.08 €
POT FRITE 2.5KG 21.05 €

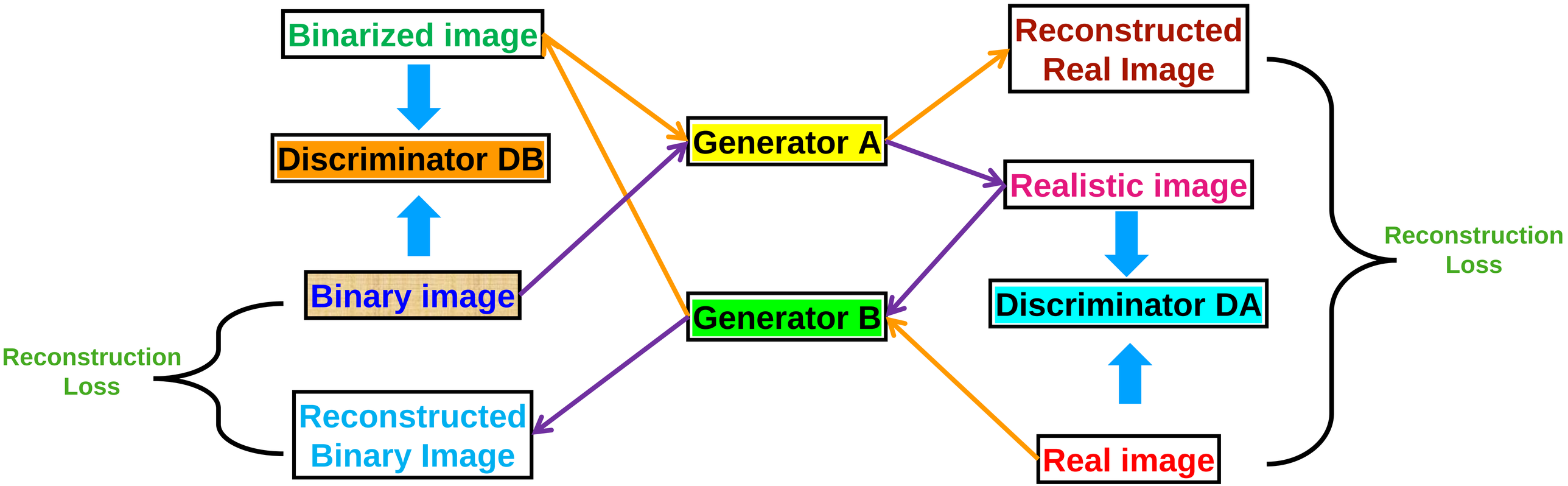


PRINCE PT DEULCE
CONFITURE.ABRICO
CX*QUICHE LORRAI
RIZ AU LAIT SAV.
MANU GOURMAI
SALADE FEUILLE D
FROMAGE GRANA PA
AUCHAN AUTHENTIQ
ASS.BOITE METAL
NETT.A/BACT.ST M
6T JAMBON SERRAN

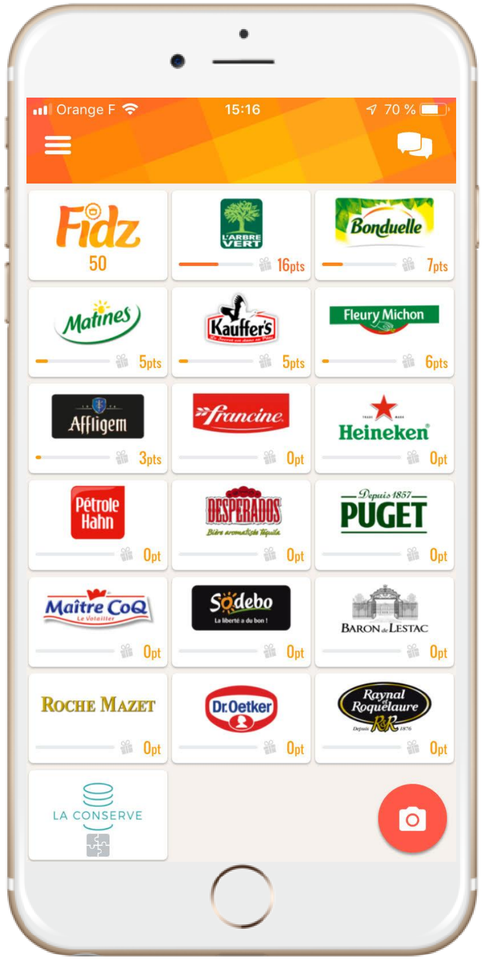


5. Synthetic image generation

DualGAN:



Agenda



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

FidMarques