Visual perception of unitary elements for layout analysis of unconstrained documents in heterogeneous databases

ICFHR 2014

September 2, 2014

Baptiste Poirriez

Aurélie Lemaitre

Bertrand Coüasnon

Irisa / Insa – Intuidoc research group

couasnon@irisa.fr

intuidoc

INSA RENNES

UMR IRISA

Context : Heterogeneous Documents

- **Document Layout Analysis**
  - Unconstrained and Heterogeneous Documents
  - Heterogeneous Databases
  - No a priori on the document

- **Maurdor Campaign dataset**
  - 2 International Competition in 2013 (www.maurdor-campaign.org)
  - Founded by the French Ministry of Defense
  - Publicly available in 2014
- **10 000 scanned documents, fully annotated**
  - Heterogeneous documents
    - Forms, Tables, Business documents, Correspondences, Faxes Diagrams, Drawings…
  - French (50%), English (25%), Arabic (25%), Mixed
  - Printed, Handwritten, Mixed
    - 1000 writers

Contrat d'entretien

N° de l'appareil: 5 3 4 2 1 0 Date de placement: .... / .... /

Nom et Prénom: COLIN Chantale

Rue: 18 rue Joseph Light N°

Code Postal:: 68290 Localité: NIEDERBRUC

Téléphone: 03 / 48 44 63 63 GSM: .... /

E-mail:

Nous marquons notre accord pour l'entretien de l'adoucisseur d'eau aux conditions suivantes:

1. Prestations annuelles

→ vérification de la vanne, du programmateur, des électrovannes et du pressostat;
→ contrôle du bon fonctionnement de l'appareil et de l'étanchéité des tuyaux de raccordement intérieurs et extérieurs;
→ nettoyage du bac à sel, du filtre d'entrée, de la gaine;
→ contrôle et réglage éventuel de la dureté.

2. Echange éventuel de pièces

Les pièces reconnues défectueuses ou usées seront remplacées et, en dehors de la garantie, facturées suivant le tarif en vigueur.

3. Paiement

Le contrat d'entretien sera payable immédiatement après la prestation au technicien qui fera l'inscription du reçu sur le bordereau d'intervention. Le contrat sera renouvelé automatiquement chaque année sauf résiliation par lettre recommandée 3 mois avant l'échéance.

DATE D'ENTRETIEN

05 / 09 / 2...
18 / 03 / 2...
10 / 10 / 2...

Fait en double exemplaire

بسم الله الرحمان الرحيم

من : القضماني إلياس
إلى : سكينة إدريسي

السلام عليكم ورحمة الله وبركاته وبعد،
يسرني في هذه اللحظة الجميلة أن أرفع قلمي هذا وأكتب
لك يا صديقتي العزيزة أسمى العبارات والتقدير التام،
وأتمنى أن تجدك رسالتي هذه في أتم الصحة وراحة بال
تام.
أختي الحبيبة: أهنئك بالمولود الجديد وأتمنى من الله العلي
القدير أن يجعله مولود خير وبركة عليك وعلى عائلتك،
ويكون زيادة في الإسلام بإذن الله.
وفي ختام رسالتي أرجو أن تسعدينا بزيارة قريبة
تغمرنا بالفرحة والسرور

والسلام

التوقيع:
إلياس

Tricycle les Bisounours pour Noël.
TOMY Aspiro Dingo pour son anniversaire le mois prochain
Peluche pour le fils d'Eleanor (naissance prévue en avril).

INVOICE

DATE: 22-2-11

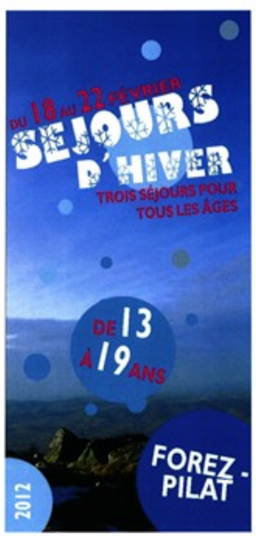
Bill To:
Ms Fiona Meery
Bankfield Road, Tyldesley
M29 8QH, Manchester

Ship To:
ANB Corporation
7 HO7 1WB
Tolsty Building
King Street, London

P.O. #	Sales Rep. Name	Ship Date	Ship Via	Terms	Due Date
A7	Elena Mor	2-3-11	Aumus Port	15 days	18-3-11

Description	Quantity	Unit Price USD	Line Total
32" LCD TV Sony Model No S-041	2 Nos	730	1460

SUBTOTAL	1460
PST 6.50%	95
GST 3.20%	47
SHIPPING & HANDLING	100
TOTAL	1702
PAID	702
TOTAL DUE	1000



ALLOCATIONS FAMILIALES 1/2 www 6008001H cerfa 11361*02

AVIS DE CHANGEMENT DE SITUATION POUR LES PRESTATIONS FAMILIALES

Livre V du code de la Sécurité Sociale

N° d'allocataire : 8357463D

Madame, Monsieur,
Si votre situation change, vos droits peuvent aussi changer. Ils peuvent augmenter ou diminuer. Vous devez donc nous prévenir de tout changement dans votre situation. Sinon, dans certains cas, vous risquez de perdre une partie de vos droits. Dans d'autres cas, vous serez obligé de nous rembourser les prestations reçues alors que vous n'y aviez plus droit.
Pour nous avertir, utilisez cet «avis de changement de situation». Complétez-le et renvoyez-le nous. **N'oubliez pas de le dater, de le signer et d'y joindre les pièces justificatives qui vous sont demandées.**

Votre caisse d'Allocations familiales.

Votre nom MARIE Votre prénom Arnaud
Votre numéro d'allocataire 8357463D

REMPLISSEZ UNIQUEMENT LES RUBRIQUES CORRESPONDANT À VOTRE NOUVELLE SITUATION

Votre nouvelle adresse
Votre adresse : 30, rue Famille ARNALDI
Code postal 06430 Commune La Brigue
Si vous avez changé de logement, précisez depuis le : 01 12 2011

Votre nouvelle situation familiale
Vous vivez en couple depuis le 03 12 2011
[X] Vous êtes mariés *(joignez une photocopie de votre livret de famille)*
[ ] Vous vivez en concubinage *(joignez une photocopie de la carte d'identité, du passeport, de la carte d'ancien combattant ou d'invalide de guerre ou encore du titre de séjour de votre concubin(e) ou partenaire)*
[ ] Vous avez signé un Pacs
Vous vivez seul(e) depuis le
[ ] Vous êtes séparé(e) de fait
[ ] Vous êtes séparé(e) légalement *(joignez une copie de la décision judiciaire)*
[ ] Vous êtes divorcé(e) *(joignez une copie du jugement de divorce)*
[ ] Vous êtes veuf(ve) *(joignez un bulletin de décès)*
[ ] Vous êtes célibataire
En cas de séparation ou de divorce, précisez qui à la garde et la charge des enfants ?

La nouvelle situation de vos enfants et des personnes à votre charge
Précisez dans le cadre ci-dessous :
• Si la situation de vos enfants ou des personnes à votre charge a changé : début d'activité professionnelle, d'apprentissage, chômage, en longue maladie, etc. ***(joignez selon la situation : bulletin de salaire, attestation du maître de stage ou d'apprentissage, certificat médical, etc.)***
• Si le nombre d'enfants ou de personnes à votre charge a changé : arrivée au foyer ***(joignez une photocopie de votre livret de famille ou de la carte d'identité ou du passeport de l'enfant)***, départ du foyer, mariage, décès...

Nom et prénom	Lien de parenté (fils, fille, neveu...) enfant recueilli	Date de naissance	Nouvelle situation	Date de ce changement de situation	Vit-il avec vous oui	non
					[ ]	[ ]
					[ ]	[ ]
					[ ]	[ ]

S. 7129c

N°493 du 17/03/2009 au 19/03/2009
Jeu multiple: 4 sur 5 matches

Matches			Cotes
10 FROSIN	- AVELLI	1	1,75
20 ARSENA	- HULLCY	1	1,30
29 2GRE	- 2MON	N	1,80
38 STETIE	- WER+1	2	1,40
49 ZEN.ST	- UDINES	1	1,70

Jeu	10	20	29	38	49	Cote
01	1	1	N	2		5,73
02	1	1	N		1	6,96
03	1	1		2	1	5,41
04	1		N	2	1	7,50
05		1	N	2	1	5,57

Mise jouée 1€
Gain potentiel max 31,2€

Vos n° hasard: de 0344 à 0348

Tirage le mardi 17/03/2009
Gain payable le lendemain du dernier match pronostiqué

MONTANT TOTAL 5€

Difficulties

- **Unconstrained and Heterogeneous Documents**
  - Methods usually need some homogeneity inside of the collection of documents
    - Stability of some blocks of text for example
- **Heterogeneous Databases**
  - Methods usually rely on document classification
  - New class for each new type of document
- **Mixing both is complex**
  - No a priori on the document
  - No a priori on the database
  - Only very general a priori on the document elements

Proposed Method : Strategy

- **Analysis of Salient Elements**
  - **Use of Perceptive Vision Mechanisms**
    - **Already used on homogeneous databases**
  - **Perceptive Vision Principle**
    - **Some contents are salient for the human vision in a document**
      - Often strongly structuring
    - **Combining several points of view**
      - Prediction/verification mechanism
      - Layout is predicted in a global vision of a document and verified with details.
  - **Multi level description of salient elements**
- **Primitives at different resolutions**
  - **Line segments**
  - **Connected components**
  - **Words recognized by OCR (Abbyy FineReader)**

Proposed Method : Strategy

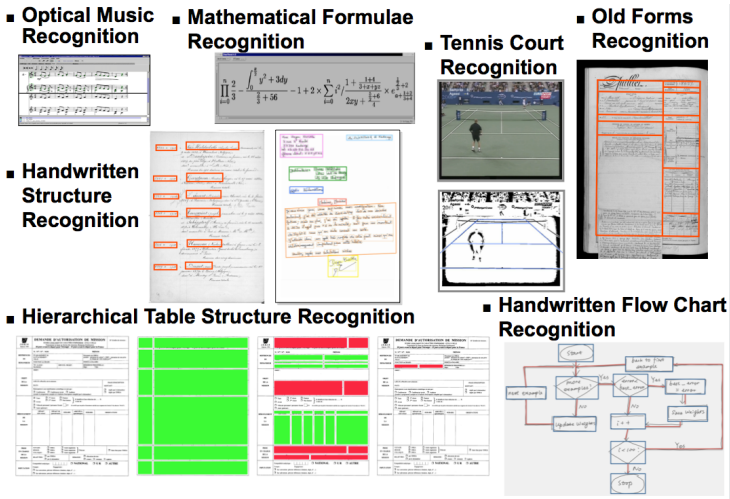
- **Detection of orientation**
  - Perceptive vision
  - Detection at low resolution of text lines (line segments)
  - Detection of the main direction of writings

- **Iteratively find the most structuring and salient elements**
  - 1: Tables, separators, boxes
  - 2: Latin printed blocks of text
  - 3: Remove of the detected salient elements : New Segmentation
  - 4: Handwritten and Arabic text lines
  - 5: Graphic regions

- **Description of each salient element**
  - General knowledge on each element

Proposed Method: Implementation with DMOS-P

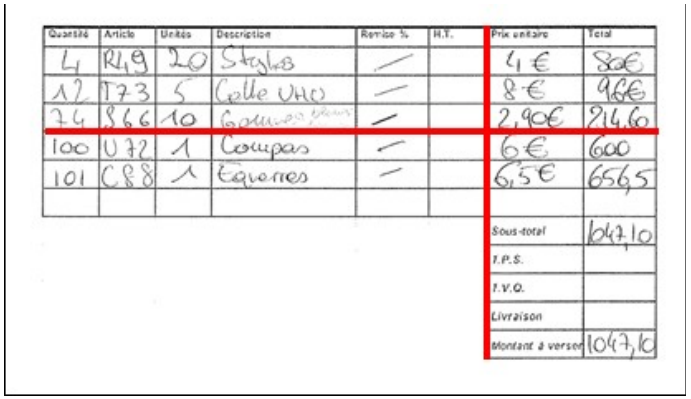
- **DMOS-P: a generic perceptive method**
  - **EPF language**
    - Grammatical description of a document
    - Generation of an adapted recognition system
  - **Validated**
    - Wide range of kind of documents
    - More than 700,000 pages processed
    - On homogeneous documents

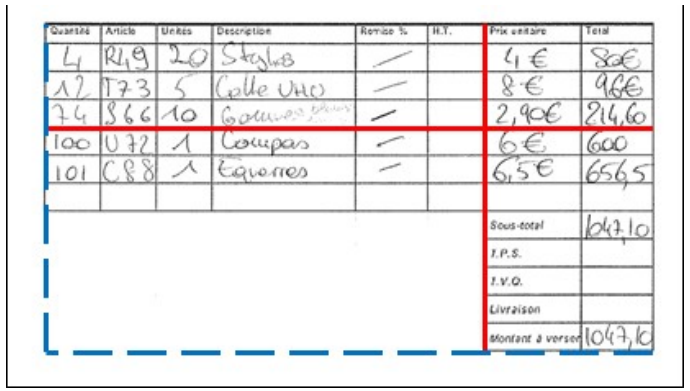


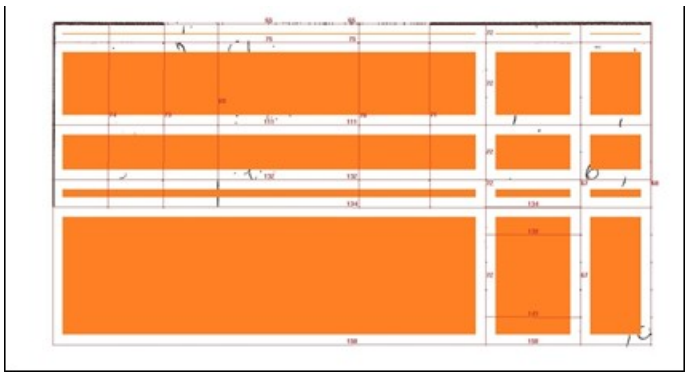
- **Application of DMOS-P for Heterogeneous Documents**
  - **Description with EPF of salient elements**
  - **New Segmentation during parsing**
  - **Combine different levels of perception**
    - Perceptive Layers
  - **Combine different salient elements**

1: Tables, separators, boxes

- **Perceptive vision**
  - Line segments at different resolution
  - Rulings detection
- **General description of a table**
  - 2 crossing rulings
  - Table frame on the cross
    - Virtual rulings
  - Set of rows
  - Set of columns
  - Recursive table detection inside each cell
- **General description of**
  - Boxes
  - Separators





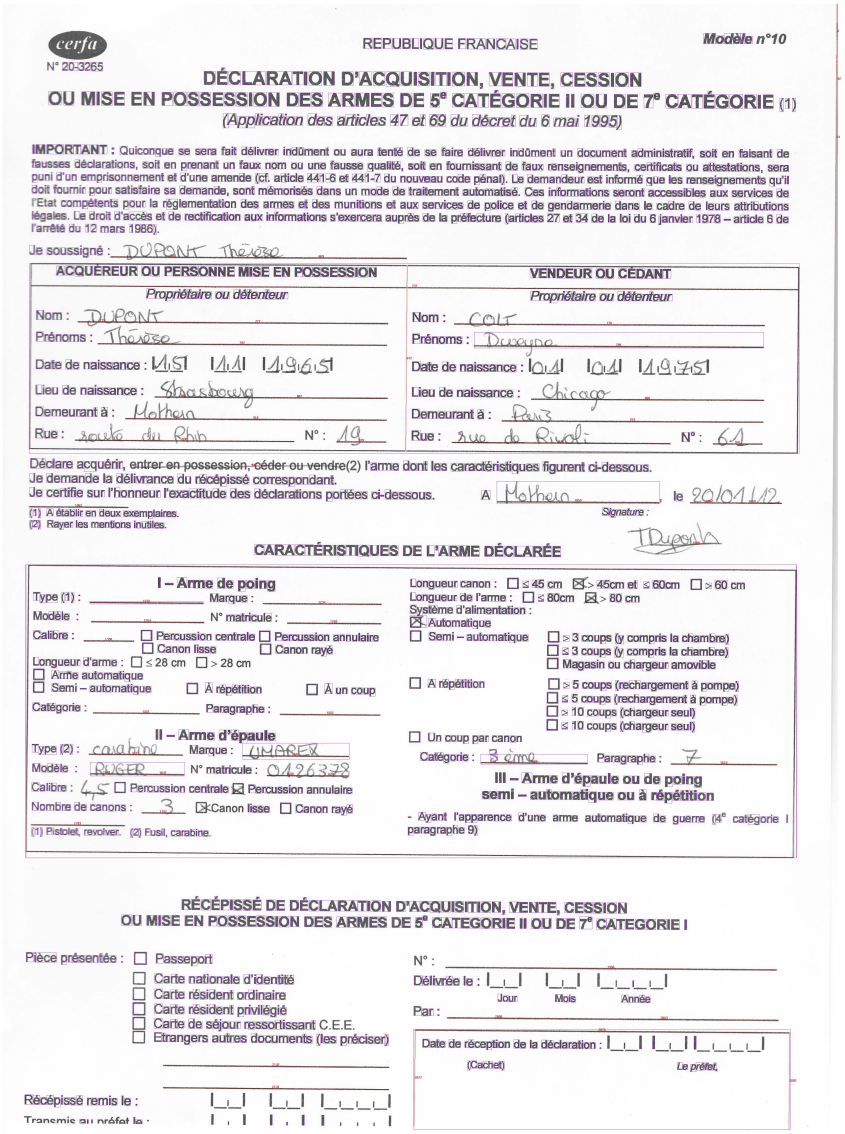


2: Latin printed block of text

- **Words detected by OCR (Abbyy Finereader)**
  - **On the complete document**
    - Errors on mixed (printed/handwritten) documents
  - **Added in a perceptive layer**

- **General description of Latin Printed block of text**
  - **Blocks built on words with a high level of confidence**
  - **Text lines and blocks**
    - With constrains on size, alignments (left, right, centered), text columns
    - Very general knowledge on blocks of text

3: New Segmentation (Remove of Table Rulings, Separators, Latin Printed Text)



cerfa
N° 20-3265

REPUBLIQUE FRANCAISE

Modèle n°10

DÉCLARATION D'ACQUISITION, VENTE, CESSION OU MISE EN POSSESSION DES ARMES DE 5e CATÉGORIE II OU DE 7e CATÉGORIE (1)

(Application des articles 47 et 69 du décret du 6 mai 1995)

IMPORTANT : Quiconque se sera fait délivrer indûment ou aura tenté de se faire délivrer indûment un document administratif, soit en faisant de fausses déclarations, soit en prenant un faux nom ou une fausse qualité, soit en fournissant de faux renseignements, certificats ou attestations, sera puni d'un emprisonnement et d'une amende (cf. article 441-6 et 441-7 du nouveau code pénal). Le demandeur est informé que les renseignements qu'il doit fournir pour satisfaire sa demande, sont mémorisés dans un mode de traitement automatisé. Ces informations seront accessibles aux services de l'Etat compétents pour la réglementation des armes et des munitions et aux services de police et de gendarmerie dans le cadre de leurs attributions légales. Le droit d'accès et de rectification aux informations s'exercera auprès de la préfecture (articles 27 et 34 de la loi du 6 janvier 1978 – article 6 de l'arrêté du 12 mars 1986).

Je soussigné : DUPONT Thérèse

ACQUÉREUR OU PERSONNE MISE EN POSSESSION *Propriétaire ou détenteur*	VENDEUR OU CÉDANT *Propriétaire ou détenteur*
Nom : DUPONT	Nom : COLT
Prénoms : Thérèse	Prénoms : Duayne
Date de naissance : 15/11/1961	Date de naissance : 01/01/1975
Lieu de naissance : Strasbourg	Lieu de naissance : Chicago
Demeurant à : Mothern	Demeurant à : Paris
Rue : route du Rhin N° : 19	Rue : rue de Rivoli N° : 61

Déclare acquérir, ~~entrer en possession, céder ou vendre~~(2) l'arme dont les caractéristiques figurent ci-dessous.
Je demande la délivrance du récépissé correspondant.
Je certifie sur l'honneur l'exactitude des déclarations portées ci-dessous. A Mothern, le 20/01/12

(1) A établir en deux exemplaires.
(2) Rayer les mentions inutiles.

Signature : TDupont

CARACTÉRISTIQUES DE L'ARME DÉCLARÉE

I – Arme de poing

Type (1) : ____ Marque : ____
Modèle : ____ N° matricule : ____
Calibre : ____ ☐ Percussion centrale ☐ Percussion annulaire ☐ Canon lisse ☐ Canon rayé
Longueur d'arme : ☐ ≤ 28 cm ☐ > 28 cm
☐ Arme automatique
☐ Semi – automatique ☐ A répétition ☐ A un coup
Catégorie : ____ Paragraphe : ____

II – Arme d'épaule

Type (2) : carabine Marque : UMAREX
Modèle : RUGER N° matricule : 01426378
Calibre : 4,5 ☐ Percussion centrale ☒ Percussion annulaire
Nombre de canons : 3 ☒ Canon lisse ☐ Canon rayé

(1) Pistolet, revolver. (2) Fusil, carabine.

Longueur canon : ☐ ≤ 45 cm ☒ > 45cm et ≤ 60cm ☐ > 60 cm
Longueur de l'arme : ☐ ≤ 80cm ☒ > 80 cm
Système d'alimentation :
☒ Automatique
☐ Semi – automatique ☐ > 3 coups (y compris la chambre) ☐ ≤ 3 coups (y compris la chambre) ☐ Magasin ou chargeur amovible
☐ A répétition ☐ > 5 coups (rechargement à pompe) ☐ ≤ 5 coups (rechargement à pompe) ☐ > 10 coups (chargeur seul) ☐ ≤ 10 coups (chargeur seul)
☐ Un coup par canon
Catégorie : 3 ème Paragraphe : 7

III – Arme d'épaule ou de poing semi – automatique ou à répétition

- Ayant l'apparence d'une arme automatique de guerre (4e catégorie I paragraphe 9)

RÉCÉPISSÉ DE DÉCLARATION D'ACQUISITION, VENTE, CESSION OU MISE EN POSSESSION DES ARMES DE 5e CATEGORIE II OU DE 7e CATEGORIE I

Pièce présentée : ☐ Passeport
☐ Carte nationale d'identité
☐ Carte résident ordinaire
☐ Carte résident privilégié
☐ Carte de séjour ressortissant C.E.E.
☐ Etrangers autres documents (les préciser)

N° : ____
Délivrée le : __/__/____ (Jour / Mois / Année)
Par : ____

Date de réception de la déclaration : __/__/____
(Cachet) Le préfet,

Récépissé remis le : __/__/____
Transmis au préfet le : __/__/____

cerfa
N°

n°10

l'Etat

DUPONT Thérèse

DUPONT Thérèse 15/11/1961 Strasbourg Mothern route du Rhin N° 19

COLT Duayne 01/01/1975 Chicago Paris rue de Rivoli 61

~~entrer en possession, céder ou vendre~~(2)

Mothern, 20/01/12

TDupont

N°

☐ ≤ 28 cm ☐ > 28 cm

carabine UMAREX RUGER N° 01426378 4,5 3 ☒ Canon

ème 7

C.E.E.

N° : __/__/____

Transmis préfet

l'arrêté du 12 mars 1986).

Je soussigné : DUPONT T

ACQUÉREUR OU PERSONNE

Propriétaire ou

Nom : DUPONT

Prénoms : Thérèse

Date de naissance : 1 1 S T 1 1

Lieu de naissance : Strasbo

Demeurant à : Mothern

Rue : route du Rhin

Déclare acquérir, entrer en possessi

Je demande la délivrance du récépi

Je certifie sur l'honneur l'exactitude

(1) A établir en deux exemplaires.

(2) Rayer les mentions inutiles.

I – Arme de

DUPONT Thé

DUPONT

Thérèse

1 1 S T 1 1 1

Strasbourg

Mothern

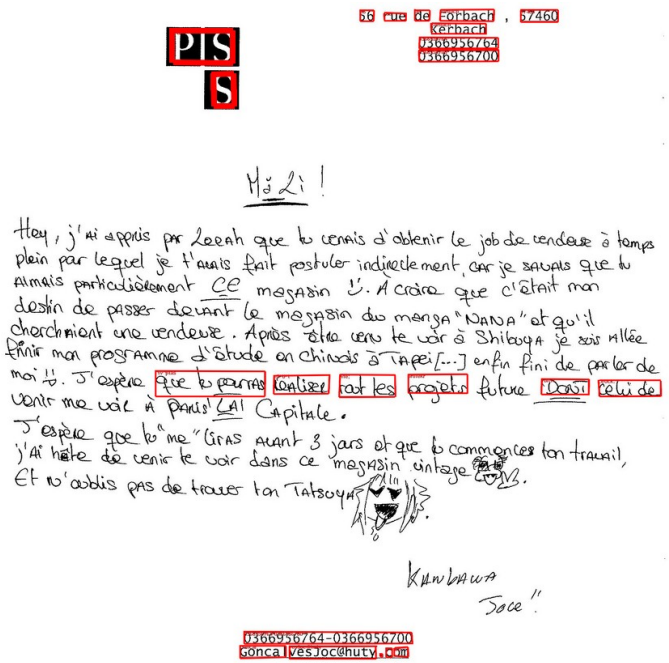
route du Rhin

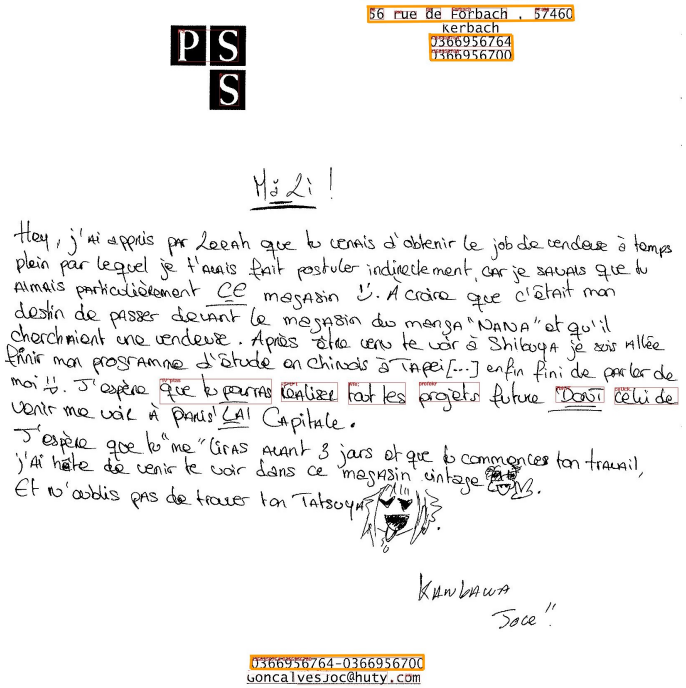
entrer en possession,

4: Handwritten and Arabic Text lines

- **New Segmentation**
  - No more tables, rulings or Printed Latin text
  - Simplify the description of handwritten and Arabic text lines
- **Perceptive Vision**
  - Text lines can be seen at low resolution as line segments
  - Confirm the text line at high resolution with regular aligned connected components
- **Description of text lines detects**
  - Handwritten text lines : Latin and Arabic
  - Printed text lines: Arabic and some remaining Latin (OCR Errors)
- **Blocks of text made of text lines**

Example of Printed/Handwritten document

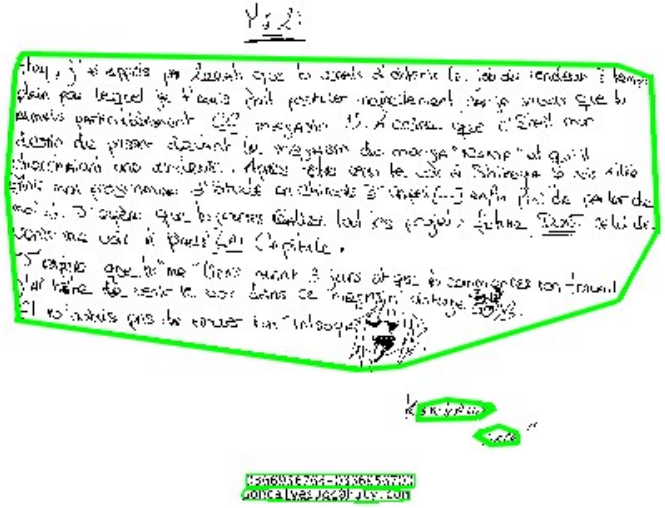




- **Words from OCR**

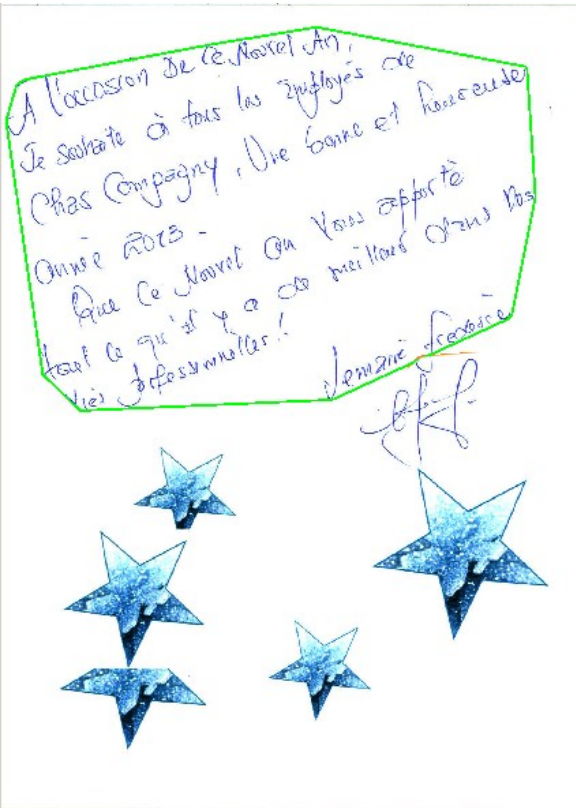
- **Latin Printed Blocks**

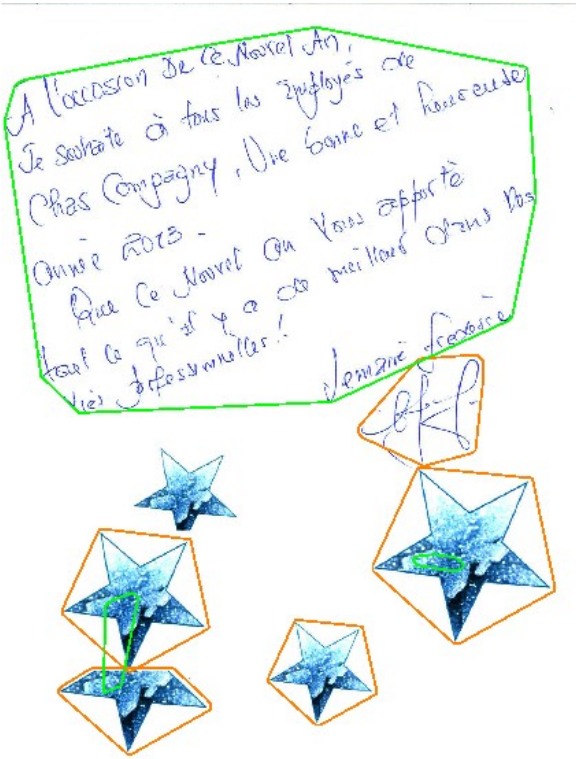
- **Blocks after Handwritten text lines detection**



5: Graphic Region

- **Simple description of Graphics**
  - **Low resolution**
  - **Remaining connected components big enough**
- **Example**





Experimental Evaluation

- **Maurdor Campaign, Module 1 (Layout Analysis task)**
  - Training set : 6,000 documents
  - Dev set : 1,000 documents
  - Test set : 1,000 documents
- **Metrics (see www.maurdor-campaign.org for details)**
  - ZoneMap (deals with split and merge) : lower better
  - Jaccard (pixel level labeling) : higher better
- **Results**
  - 2nd on ZoneMap
  - Close to best on Jaccard

Participant	ZoneMap(%)	Jaccard
Participant 1	**48.7**	**0.45**
Our method	*59.2*	*0.44*
Participant 2	73.5	0.28

- **Jaccard by class**

Participant	Text zone	Graphic zone	Table
Participant 1	0.552	0.394	**0.363**
Our method	***0.553***	***0.402***	*0.307*
Participant 2	0.307	0.176	0.174

Conclusion

- **DLA on Heterogeneous and Unconstrained Documents**
  - Difficult task, open problem
  - General grammatical description of document elements
    - Perceptive vision mechanism
  - Iterative recognition of salient elements
  - Ability to re-segment the document during the analysis
  - New application of DMOS-P
    - Up to now applied on homogeneous databases
- **Improvements**
  - Too much split and merge
    - Need more homogeneous blocks
    - Avoid confusion between graphics and text
  - Introduce classifiers for adding local information
    - Printed/handwritten detection
    - Language detection
    - Kind of graphics detection
  - Mixing those classifiers with grammatical description