

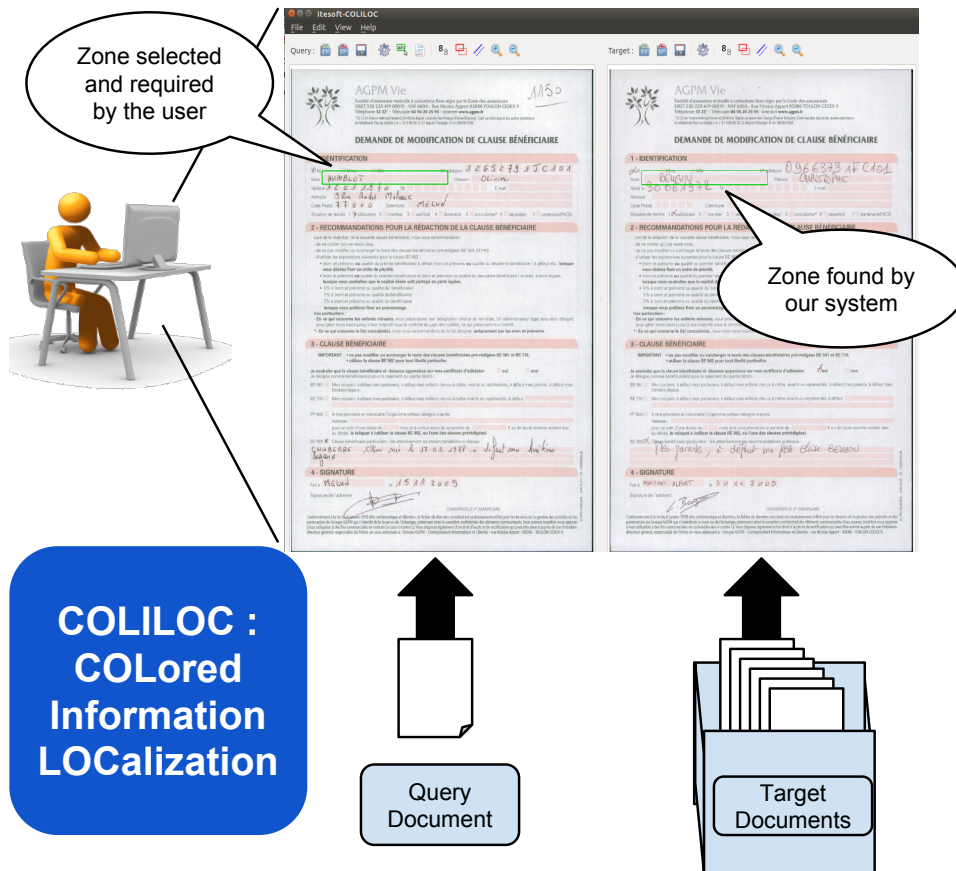
Minimum Cost Subgraph Matching Using a Binary
Linear Program
Application sur les documents

Julien LEROUGE
Maroua HAMMAMI
Sébastien ADAM , Pierre HÉROUX

September 29, 2015

- ① Introduction
- ② Extraction des zones informatives
- ③ Construction de graphes
- ④ Recherche des sous graphes
- ⑤ Résultats
- ⑥ Conclusion

COLILOC : recherche du texte dans images de documents administratifs et commerciaux



Objectifs

- Extraction de zones informatives

Contraintes

- peu d'exemples d'apprentissage

Problématiques

- Variabilité des images (numérisation, contenu du document...)

localisation à base de contenus

- représentation structurelle
- Tolerant subgraph isomorphism

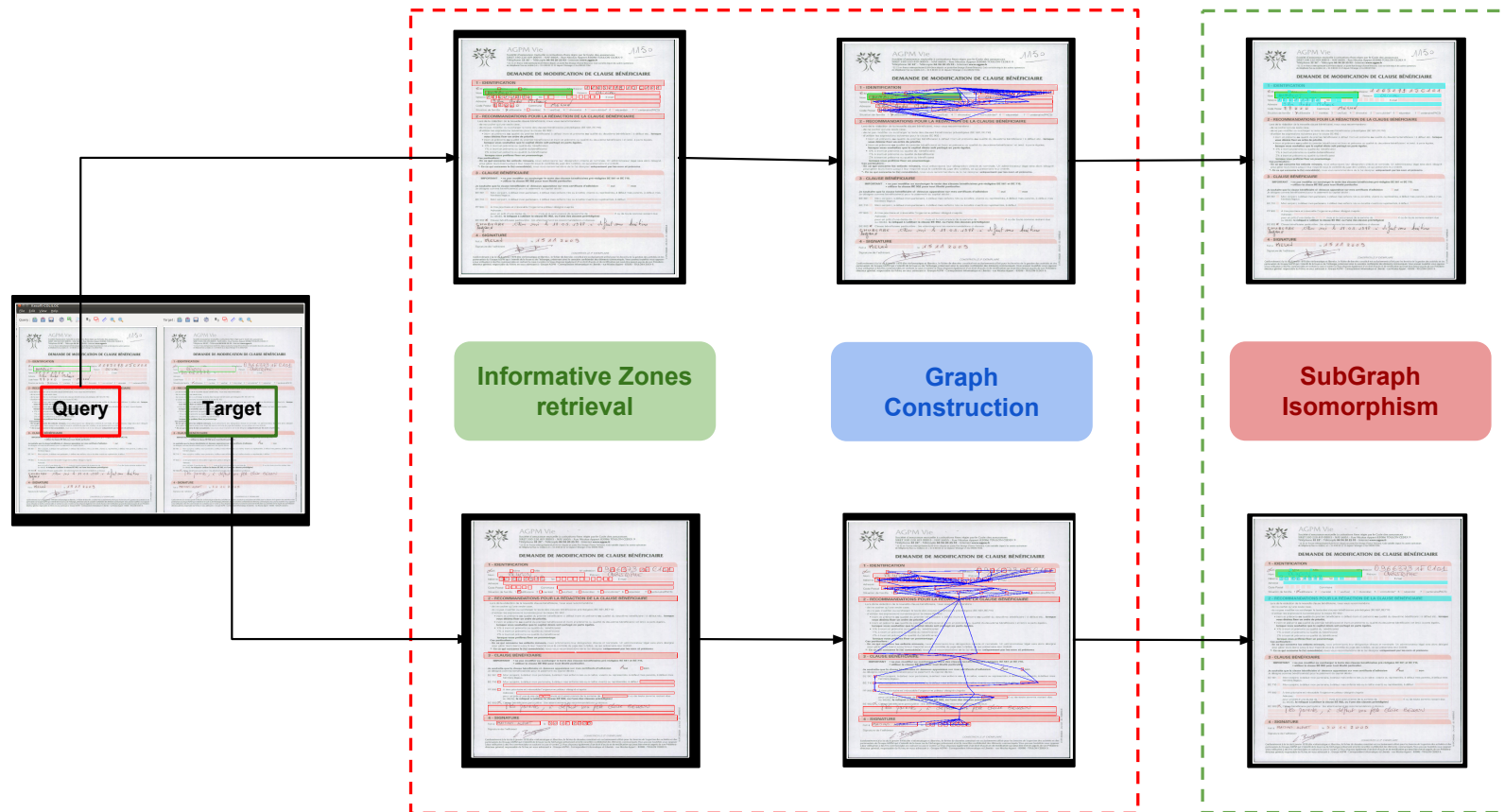
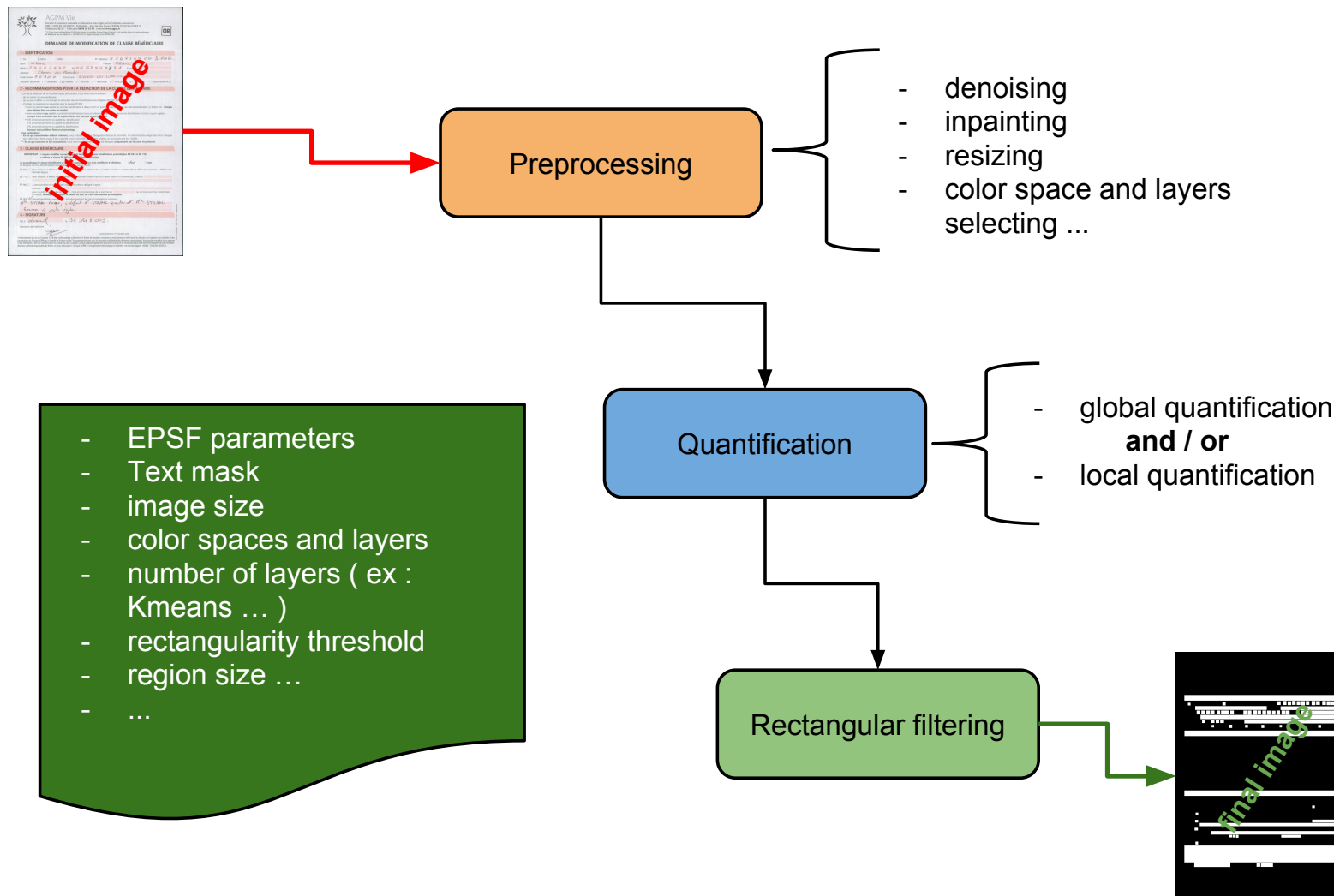


Figure : Processus de localisation du texte

Approche (1) : Extraction des zones informatives



Approche (2) : Construction du graphe

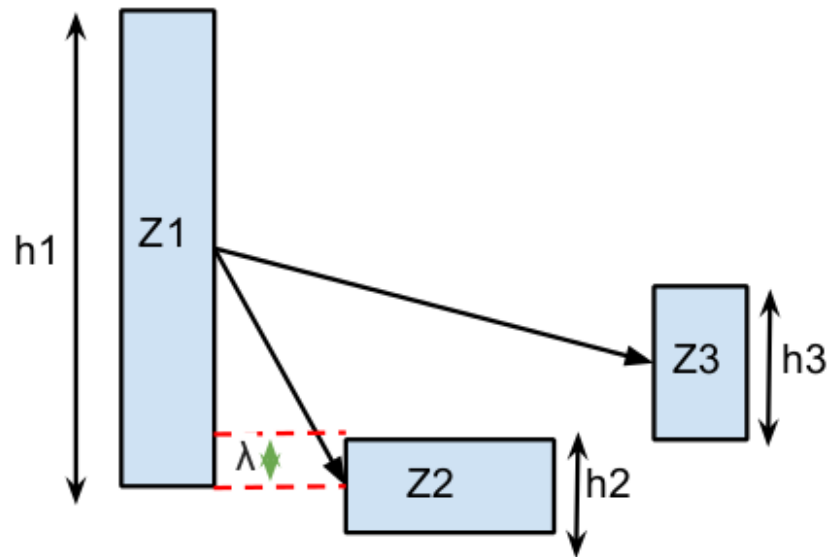


Figure : visibilité verticale

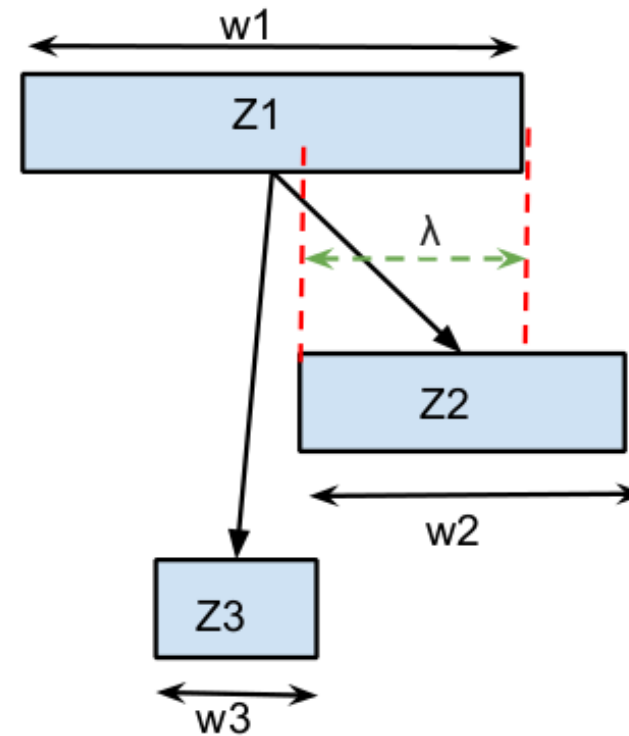


Figure : visibilité horizontale

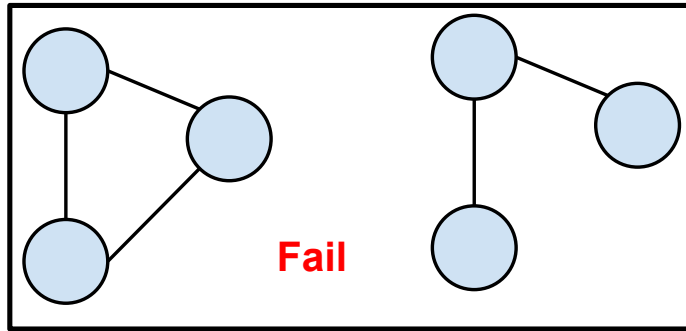
Graphes distordus

- zones manquantes , fusionnées , divisées ...
- Arcs en plus ou en moins

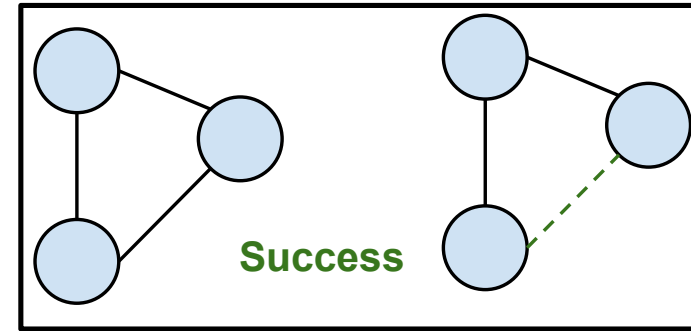
Solution : Minimum cost Subgraph matching (MCSM)

- Tolérant aux erreurs topologiques
- Programmation linéaire de la distance d'édition
- Distance d'édition
 - Substitution : Noeuds / Arcs
 - Création : Noeuds / Arcs

Minimum Cost Subgraph Matching (MCSM) : nouvelle formulation



Substitution-Tolerant Subgraph Matching (STSM)



Minimum Cost Subgraph Matching (MCSM)

Formulation : MCSM

$$\min_{\mathbf{x}, \mathbf{y}, \alpha, \beta} \left(\sum_{i \in \mathcal{V}_1} \sum_{k \in \mathcal{V}_2} x_{i,k} \cdot c(i \rightarrow k) + \sum_{i \in \mathcal{V}_1} \alpha_i \cdot c(i \rightarrow \epsilon) + \sum_{ij \in \mathcal{E}_1} \sum_{kl \in \mathcal{E}_2} y_{ij,kl} \cdot c(ij \rightarrow kl) + \sum_{ij \in \mathcal{E}_1} \beta_{ij} \cdot c(ij \rightarrow \epsilon) \right) \quad (1)$$

BLP : fonction objectif

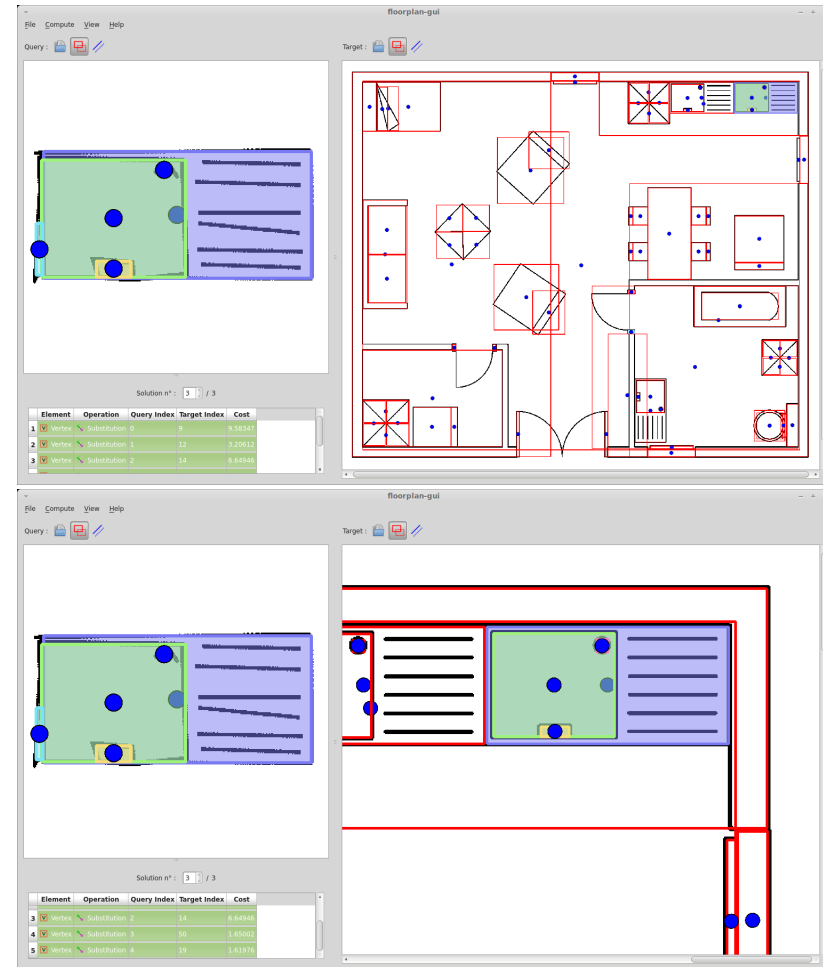
$$\min_{\mathbf{x}, \mathbf{y}} \left(\sum_{i \in \mathcal{V}_1} \sum_{k \in \mathcal{V}_2} x_{i,k} \cdot (c(i \rightarrow k) - c(i \rightarrow \epsilon)) + \sum_{i \in \mathcal{V}_1} c(i \rightarrow \epsilon) \right. \\ \left. + \sum_{ij \in \mathcal{E}_1} \sum_{kl \in \mathcal{E}_2} y_{ij,kl} \cdot (c(ij \rightarrow kl) - c(ij \rightarrow \epsilon)) + \sum_{ij \in \mathcal{E}_1} c(ij \rightarrow \epsilon) \right) \quad (2)$$

$$\text{S.t. : } \begin{aligned} & \sum_{k \in \mathcal{V}_2} x_{i,k} \leq 1 \forall i \in \mathcal{V}_1 \\ & \sum_{i \in \mathcal{V}_1} x_{i,k} \leq 1 \forall k \in \mathcal{V}_2 \\ & \sum_{l \in \mathcal{V}_2, kl \in \mathcal{E}_2} y_{ij,kl} \leq x_{i,k} \forall ij \in \mathcal{E}_1, \forall k \in \mathcal{V}_2 \\ & \sum_{k \in \mathcal{V}_2, kl \in \mathcal{E}_2} y_{ij,kl} \leq x_{j,l} \forall ij \in \mathcal{E}_1, \forall l \in \mathcal{V}_2 \\ & x_{i,k} \in \{0, 1\} \forall i \in \mathcal{V}_1, \forall k \in \mathcal{V}_2 \\ & y_{ij,kl} \in \{0, 1\} \forall ij \in \mathcal{E}_1, \forall kl \in \mathcal{E}_2 \end{aligned}$$

Résultats (1) : Base Floorplan



STSM



MCSM

Résultats (2) : Base Floorplan

Method	MCSM					STSM
rC	5	10	20	40	80	-
0	0.95	0.99	1.00	1.00	1.00	0.99
4	0.75	0.92	0.94	0.94	0.94	0.93
8	0.58	0.80	0.84	0.85	0.84	-

Table : Mean F1-score on nodes matching rate for the 5500 queries of the dataset

Method	MCSM					STSM
rC	5	10	20	40	80	-
0	0.11	0.13	0.17	0.41	3.21	3.09
4	0.16	0.13	0.19	0.52	4.72	4.92
8	0.22	0.18	0.27	1.14	10.05	9.26

Table : Mean elapsed time by correctly found instance in seconds

Expériences et Résultats (1) : Base de documents COLILOC

Protocole d'évaluation

- 130 documents (8 classes)
- 50 graphes requêtes par classe

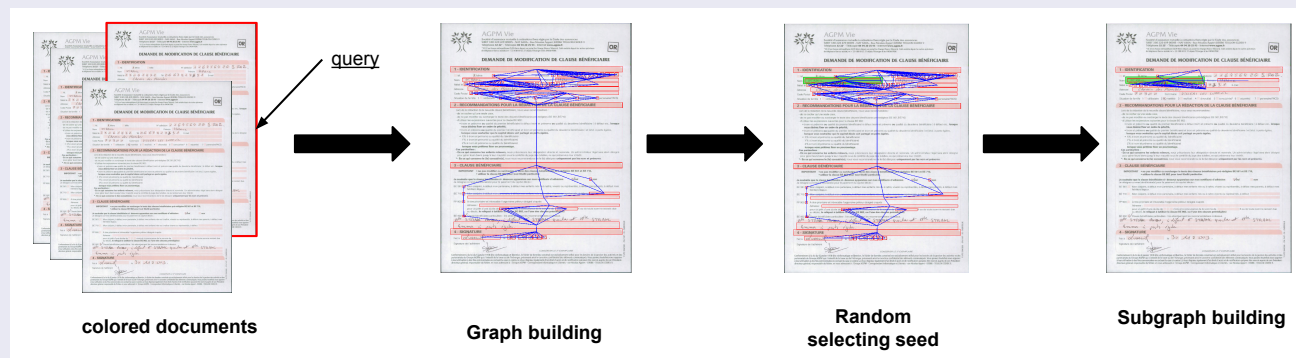


Figure : Processus de génération des graphes requêtes

Expériences et Résultats (2) : Base de documents COLILOC

Paramètres à étudier

- Prétraitement
 - EPSF
 - inpaint
- Quantification de couleurs : *Kmeans*
 - K couches
 - Espace couleur

Inpaint	Yes / No
EPSF	Yes / No
Color space	$\in \{RGB, YIQ, Lab\}$
K	$\in \{2, 3, 4\}$

36 configurations \Rightarrow 36 différents graphes

Expériences et Résultats (2) : Base de documents COLILOC

Résultats

		Meilleures configurations				Meilleur résultat
		Prétraitement		Quantification		
		EPSF	Inpaint	K	Espace couleur	
Base		non	oui	2	<i>YIQ</i>	86.9%
Classe	1 (11)	oui	oui	2	<i>YIQ</i>	83.45%
	2 (31)	non	oui	2	<i>YIQ</i>	90.38%
	3 (29)	oui	oui	2	<i>Lab</i>	93.44%
	4 (14)	oui	oui	2	<i>YIQ</i>	99.28%
	5 (12)	oui	non	3	<i>RGB</i>	100%
	6 (11)	non	oui	2	<i>YIQ</i>	83.81%
	7 (9)	oui	oui	2	<i>YIQ</i>	92.66%
	8 (13)	non	oui	4	<i>RGB</i>	72%

- Localisation de texte dans des images de documents
 - ① Extraction des zones informatives
 - ② Construction de graphes
 - ③ Recherche de sous graphes
- L'isomorphisme des sous graphes : (MSCM)
 - Tolérant aux erreurs topologiques
 - Formulation en programmation linéaire de la distance d'édition