



Filtrage de descripteurs locaux pour l'amélioration de la détection de documents

Marçal Rusiñol, Joseph Chazalon, Jean-Marc Ogier

► To cite this version:

Marçal Rusiñol, Joseph Chazalon, Jean-Marc Ogier. Filtrage de descripteurs locaux pour l'amélioration de la détection de documents. Colloque International Francophone sur l'Écrit et le Document 2016 (CIFED), Mar 2016, Toulouse, France. pp.589-604. hal-01337266

HAL Id: hal-01337266

<https://hal.science/hal-01337266>

Submitted on 24 Jun 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Filtrage de descripteurs locaux pour l'amélioration de la détection de documents¹

Marçal Rusiñol^{*} — Joseph Chazalon[◇] — Jean-Marc Ogier[◇]

^{*} *Computer Vision Center, Univ. Autònoma de Barcelona, España*

[◇] *L3i, Université de La Rochelle, France*

RÉSUMÉ. *Cet article propose une méthode simple et efficace qui vise à réduire la quantité de descripteurs locaux à indexer dans un scénario de mise en correspondance d'images de documents. Au cours d'une étape d'entraînement préalable, la mise en correspondance entre un document modèle et un ensemble d'images contenant une instance de ce modèle est calculée en ne retenant que les descripteurs locaux du modèle qui produisent des appariements corrects de façon régulière. Cette approche a été évaluée sur la base de données ICDAR2015 SmartDOC qui contient environ 25 000 images de documents capturées avec un dispositif mobile. L'intérêt de cette étape de filtrage a été testée pour les descripteurs ORB et SIFT. Les résultats montrent un gain tant au niveau de la qualité des appariements finaux que du temps et de l'espace nécessaires aux calculs.*

ABSTRACT. *In this paper we propose an effective method aimed at reducing the amount of local descriptors to be indexed in a document matching framework. In an off-line training stage, the matching between the model document and incoming images is computed retaining the local descriptors from the model that steadily produce good matches. We have evaluated this approach by using the ICDAR2015 SmartDOC dataset containing near 25 000 images from documents to be captured by a mobile device. We have tested the performance of this filtering step by using ORB and SIFT local detectors and descriptors. The results show an important gain both in quality of the final matching as well as in time and space requirements.*

MOTS-CLÉS : *descripteurs locaux, capture mobile, détection de documents, sélection de points d'intérêt.*

KEYWORDS: *local descriptors, mobile capture, document matching, keypoint selection.*

1. Traduction d'un article publié en anglais à CBDAR 2015 (Chazalon *et al.*, 2015b)

1. Introduction

La mise en correspondance d'images de documents en situation nomade, à l'aide de dispositifs mobiles, a suscité l'intérêt de notre communauté au cours des dernières années. Elle consiste en 2 fonctions, qui peuvent être utilisées séparément ou conjointement : l'identification d'une instance de document (parmi un ensemble prédéfini de modèles) dans une image capturée en mobilité (trame vidéo ou photo), et l'estimation de la transformation du document modèle vers la scène capturée. Plusieurs applications dérivent de cette approche, comme la capture mobile guidée par modèle (Rusiñol *et al.*, 2015), la réalité augmentée sans marqueurs, la fouille de documents (Nakai *et al.*, 2005b) ou le *stitching* et *mosaicking* de documents (Luqman *et al.*, 2013). Les techniques proposées dans la communauté de l'analyse de documents sont inspirées de techniques classiques de reconnaissance d'objets, et un effort particulier a été consacré à la mise en place des descripteurs locaux performants adaptés au problème spécifique de la mise en correspondance de documents.

Ainsi, Nakai et al (Nakai *et al.*, 2005b ; Nakai *et al.*, 2005a) ont proposé la méthode de *Locally Likely Arrangement Hashing* (LLAH) qui encode la distribution géométrique des mots en calculant un ensemble de descripteurs invariants aux effets de perspective. Dans (Moralada et Hull, 2010), les auteurs proposent une approche similaire qui capture également la distribution spatiale d'éléments voisins.

Cependant, malgré l'importance de la recherche de nouveaux descripteurs locaux performants face aux images de documents, il est pertinent de se préoccuper de l'efficacité de ces systèmes, en particulier dans l'optique d'une exploitation sur périphérique mobile. Dans le cas de LLAH par exemple, la capacité de cette méthode à répondre à des requêtes en quelques millisecondes sur une base indexée de millions de pages se fait au prix d'une occupation mémoire importante pour le stockage des descripteurs, ce qui la rend inutilisable avec les périphériques mobiles actuels.

Des descripteurs binaires comme BRIEF (Calonder *et al.*, 2012) ou ORB (Rublee *et al.*, 2011), en combinaison avec des méthodes d'indexation comme *Locality Sensitive Hashing* (Andoni et Indyk, 2008) (LSH), permettent quant à eux la mise en correspondance d'images de documents en temps réel sur dispositifs mobiles. Toutefois, ces descripteurs binaires ont un pouvoir de discrimination plus faible que d'autres descripteurs de l'état de l'art comme SIFT (Lowe, 2004) ou SURF (Bay *et al.*, 2008). De plus, même si LSH permet de calculer des mises en correspondance directes en temps réel, cette méthode montre des difficultés de passage à l'échelle.

Un élément critique de ces systèmes est donc la quantité de mémoire nécessaire, due à l'énorme quantité de points d'intérêt qui sont extraits. Typiquement, les paramètres par défauts des méthodes citées, lorsque l'on cherche à extraire les points d'intérêt d'images de documents acquises à 300 DPI, entraînent la génération de milliers d'éléments. Il est évidemment possible d'ajuster les paramètres des détecteurs (seuil de confiance, etc.) pour limiter le nombre de points extraits. Cependant, cette réduction quantitative provoque une réduction qualitative car il n'est pas possible à cette étape de garantir le pouvoir discriminant des descripteurs retenus ; la réponse

du détecteur est le seul critère disponible. Il est en effet tout à fait possible que les points d'intérêt présentant la meilleure réponse pour le détecteur (courbure, contraste, etc.) ne présentent que peu d'intérêt au niveau de leur description (peu discriminante). C'est typiquement le cas dans les images de documents où beaucoup de motifs locaux se répètent, dans les zones de texte en particulier.

Cet article propose une méthode simple mais cependant efficace qui vise à réduire la quantité de descripteurs locaux à indexer dans un système de mise en correspondance de documents. Inspirés par le travail de Kurz et al. (Kurz *et al.*, 2012), nous proposons d'apprendre quels sont les points d'intérêt les plus pertinents pour un modèle donné à partir d'un ensemble d'apprentissage composé d'exemples d'apparitions du modèle dans différentes scènes. En mettant en correspondance le modèle et chaque image de la séquence d'apprentissage, il est possible de déterminer quels sont les points d'intérêt dont les descripteurs permettent des appariements robustes. Cette approche a été évaluée sur la base de données de la compétition SmartDOC d'ICDAR 2015, qui contient environ 25 000 images de documents capturées à l'aide d'un *smart-phone*. L'intérêt de notre approche a été testée pour les descripteurs SIFT et ORB. Les résultats montrent un gain tant au niveau de la qualité des appariements finaux que du temps de calcul et de l'occupation mémoire.

L'article est organisé de la façon suivante : la section 2 introduit brièvement l'algorithme de mise en correspondance exploitant les descripteurs locaux. La section 3 présente la méthode de filtrage proposée. La section 4 présente notre protocole expérimental et la section 5 les résultats obtenus. Finalement, la section 6 présente nos conclusions.

2. Détection de documents

Nous avons suivi une approche usuelle de mise en correspondance de documents utilisant des descripteurs locaux. Pour chaque document d_i d'un ensemble de modèles à indexer $D = \{d_1, d_2, \dots, d_M\}$, nous extrayons un ensemble $K_i = \{k_1, k_2, \dots, k_N\}$ de N points d'intérêt et calculons les descripteurs locaux associés. Dans le cas présenté dans cet article, les descripteurs SIFT et ORB obtenus sont invariants à la rotation et aux changements d'échelle, d'illumination et de perspective. Chaque point d'intérêt k_j est alors décrit par un vecteur de caractéristiques f_j (SIFT ou ORB). Ces descripteurs sont alors indexés à l'aide l'architecture FLANN (Muja et Lowe, 2009) qui utilise des *KD-trees* pour les descripteurs SIFT ou *LSH* (Andoni et Indyk, 2008) pour les descripteurs binaires de ORB.

Pour chaque image requête du dispositif mobile, des points d'intérêt sont détectés et leurs descripteurs sont calculés, puis mis en correspondance avec la table d'indexation. Pour pouvoir produire des mises en correspondance stables, nous utilisons le *ratio-test* proposé par Lowe (Lowe, 2004), qui considère qu'un appariement est acceptable dès lors que le rapport entre la distance au premier et au second voisin (dans l'espace des caractéristiques) dépasse un certain seuil.

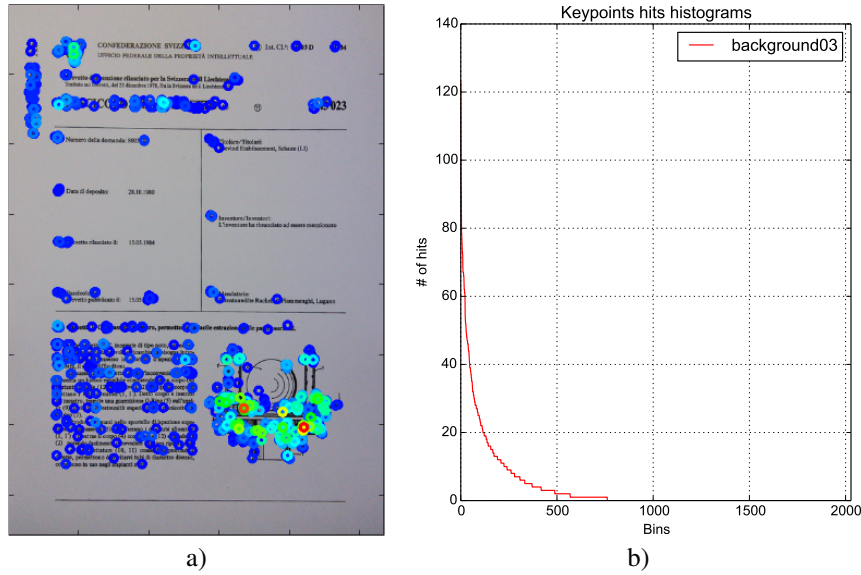


Figure 1. Importance relative des descripteurs locaux lors de la mise en correspondance d'une image modèle et celles d'une séquence de test : très peu de points d'intérêt servent effectivement de support à l'estimation de l'homographie. a) Points d'intérêt extraits (le bleu indique les points peu utilisés, le cyan et le vert ceux fréquemment utilisés, le orange et le rouge ceux très fréquemment utilisés), b) histogramme d'utilisation de ces points après validation géométrique (RANSAC).

À partir de cet ensemble d'appariement potentiels, une étape de validation géométriques basée sur la méthode *RANSAC* (Fischler et Bolles, 1981) est utilisée pour filtrer les appariements aberrants et pour estimer l'homographie entre le modèle reconnu et son instance visible dans la scène.

3. Filtrage de descripteurs locaux

La section précédente a montré l'importance des points d'intérêt extrait de chaque modèle dans la qualité des mises en correspondances qui peuvent être réalisées avec la méthode usuelle. Cette approche présente à cet égard deux points faibles. D'une part, les heuristiques permettant l'extraction de points d'intérêt se basent uniquement sur des critères locaux tels que le contraste ou la courbure, et ne garantissent donc pas le caractère discriminant des descripteurs associés. L'étape de sélection ne permet donc que d'extraire des éléments robustes aux déformations, sans considérer leur caractère discriminant. D'autre part, l'étape de détection produit généralement un grand nombre de points pour lesquels les descripteurs sont ambigus, et seront éliminés au cours des étapes successives d'appariement (test de stabilité, ou validation géométrique).

Ces deux problèmes sont bien illustrés par la figure 1. La figure 1a) montre un document modèle et les points d'intérêt qui en ont été extraits, représentés par des cercles. Cette illustration montre la fréquence d'utilisation de chaque point avec une échelle de couleur : le bleu et le vert indiquent les points qui sont rarement mis en correspondance tandis que des couleurs oranges et rouges indiquent ceux qui sont les meilleurs supports pour l'estimation de la transformation de perspective entre le modèle et les images capturées en mobilité. On remarque la forte présence de points sur les zones textuelles, dont les descripteurs sont ambigus (d'où leur faible utilité pour l'appariement). De plus, on constate qu'une part importante des points d'intérêt extraits par le détecteur sont inutiles lors de la mise en correspondance. Ceci est confirmé par la figure 1b) : l'histogramme d'utilisation des points d'intérêt pour l'estimation la transformation de perspective, c'est à dire les points qui n'ont pas été filtrés par le *ratio-test* ni par *RANSAC*, montre une diminution très rapide de l'importance des points extraits.

Afin d'améliorer la robustesse des points d'intérêt en estimant le caractère discriminant de leurs descripteurs, nous avons suivi une approche inspirée du travail de Kurz et al. (Kurz *et al.*, 2012). Nous proposons de construire des séquences d'apprentissage qui contiennent des représentations du document modèle à indexer. Dans ces séquences, la mise en correspondance entre le modèle et les images est calculée pour ne retenir que les descripteurs locaux qui produisent de bons appariements de façon stable.

Plus concrètement, les étapes de la méthode sont les suivantes, étant donnés un document modèle et une séquence d'apprentissage formée d'image contenant une instance du modèle dans une scène :

- 1) extraire une quantité importante de points d'intérêt du document modèle et calculer les descripteurs associés ;
- 2) construire un histogramme d'utilisation effective des descripteurs (et des points d'intérêt associés) lors de l'appariement entre le modèle et les instance dans les images d'apprentissage (l'estimation de l'homographie), après validation géométrique ;
- 3) construire un nouvel ensemble de points et descripteurs en ne conservant que ceux qui ont exhibé les meilleurs réponses ;
- 4) utiliser l'ensemble de points et descripteurs filtrés dans un système amélioré.

Il faut remarquer que cette séquence d'apprentissage ne requiert pas de connaître la position précise du document dans chaque image ; il suffit de savoir, pour une séquence d'apprentissage, quel document modèle est visé. Les étapes de filtrage permettent de contrôler la cohérence intrinsèque de la réponse et de sélectionner les meilleurs descripteurs locaux.

4. Protocole expérimental

Cette section présente la base de données, le protocole d'évaluation, et les mesures utilisées pour évaluer le gain en performance apporté par le filtrage proposé.

4.1. Base de données

Notre base de données de test est la base SmartDOC pour la capture de documents en mobilité (challenge 1) (Burie *et al.*, 2015). Elle consiste en six types de documents différents provenant de bases publiques, chaque type contenant 5 documents. Ces types ont été retenus afin de couvrir différentes formes d'organisation spatiale et différents contenus : complètement textuels ou ayant une grande quantité d'information graphique. Ces documents proviennent de documents techniques et des brevets de la base Ghenga (Medvet *et al.*, 2011), des premières pages d'articles scientifiques de la base MARG (Ford et Thoma, 2003), des premières pages de magazines en couleur du *PRIMA layout analysis dataset* (Antonacopoulos *et al.*, 2009), des formulaires d'imposition du *NIST Tax Forms Dataset* (SPDB2) (Dimmick *et al.*, 1991), et finalement des lettres dactylographiées de la base Tobacco800 (Lewis *et al.*, 2006). Quelques exemples d'images de cette base sont visibles à la figure 2.

Chacun de ces documents a été imprimé à l'aide d'une imprimante laser couleur, puis capturé avec un *smartphone* Google Nexus 7. Des séquences vidéo courtes (environ 10 secondes) ont été acquises pour chacun des 30 documents en utilisant cinq fonds différents. Le tableau 1 et la figure 3 illustrent plus en détail les fonds et les conditions de capture mentionnés. Les séquences ont été capturées avec une résolution *Full HD* (1920x1080) et un *frame rate* variable. Puisqu'elles ont été acquises en tenant le dispositif à la main, ces séquences présentent des distorsions réalistes comme du flou optique ou de mouvement, des variations d'illumination et même des occlusions des documents. En résumé, la base consiste en 150 clips vidéo contenant au total près de 25 000 trames. En plus de ces vidéos, des photos en résolution 8 MPixels de chaque document ont été acquises et utilisées comme modèles.

Cette collection a été annotée pour produire une vérité terrain des coordonnées du document visibles sur chaque trame (les détails de cette approche sont présentés dans (Chazalon *et al.*, 2015a ; Burie *et al.*, 2015)).

4.2. Protocole d'évaluation

L'évaluation conduite compare le système classique de filtrage de points d'intérêt, basé sur des heuristiques liées aux paramètres du détecteur (*baseline*), et le système de filtrage proposé (*optimized*). Nous décrivons ici comment les résultats ont été obtenus pour chaque système. Dans ce qui suit, on veillera à distinguer K_i^x , l'ensemble formé des x points extraits par le détecteur (ORB ou SIFT) dans le modèle d_i , de $K_i^x \setminus y$, l'ensemble des y points conservés après filtrage dans l'ensemble précédent.

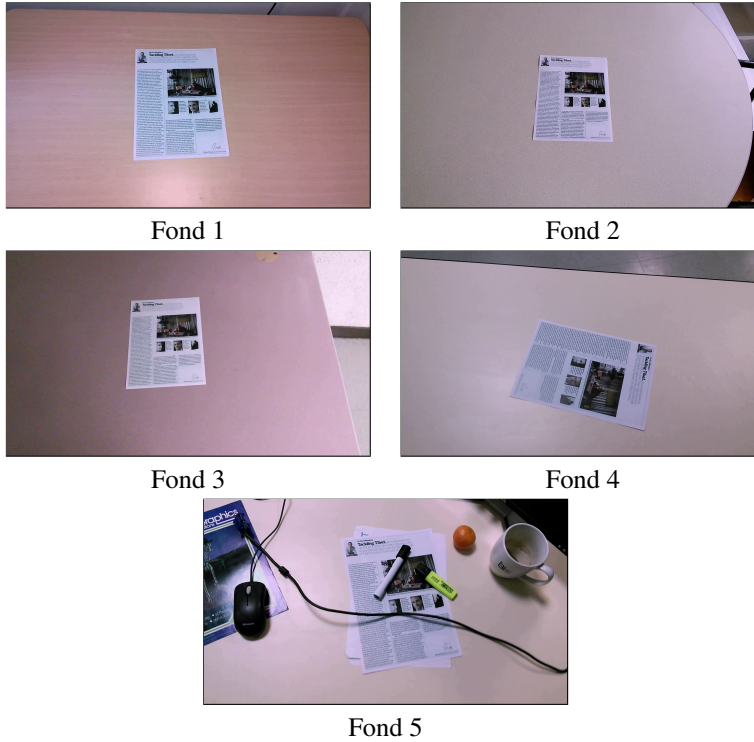


Figure 3. Les différents fonds de la base SmartDOC.

4.2.1. Référence

La mise en correspondance de documents en utilisant des descripteurs locaux avec les paramètres par défaut est mise en place avec l'algorithme suivant, où D est l'ensemble de documents modèles, T est l'ensemble de différents nombre de points d'intérêt à extraire, S_i est l'ensemble de j séquences vidéos (une pour chaque fond) pour chaque document modèle d_i .

```

pour chaque document  $d_i \in D$  faire
  pour chaque seuil  $t \in T$  faire
    construire  $K_i^t$ , l'ensemble de  $t$  points extraits de  $d_i$ 
      avec les paramètres par défaut
    pour chaque séquence  $s_{i,j} \in S_i$  faire
      traiter  $s_{i,j}$ , apparier chaque trame avec  $d_i$  en utilisant  $K_i^t$ 
    fin pour
  fin pour
fin pour

```

4.2.2. Filtrage

La mise en correspondance de documents en utilisant des descripteurs locaux sélectionnés par la méthode de filtrage est mise en place avec l'algorithme suivant, où D est l'ensemble de documents modèles, T est l'ensemble de différents nombre de points d'intérêt à conserver, S_i est l'ensemble de j séquences vidéos (une pour chaque fond) pour chaque document modèle d_i .

```
pour chaque document  $d_i \in D$  faire
  extraire  $K_i^{init}$ , l'ensemble contenant  $init$  points de  $d_i$  avec les paramètres par défaut
  pour chaque séquence d'apprentissage  $s_{i,j} \in S_i$  faire
    construire  $h_{i,j}$ , l'histogramme de fréquences d'utilisation de  $K_i^{init}$ 
    pour chaque trame de  $s_{i,j}$ 
      pour chaque seuil  $t \in T$  faire
        construire  $K_i^{init} \setminus t$ , l'ensemble des  $t$  meilleurs points de  $h_{i,j}$ 
        pour chaque séquence de test  $s_{i,t} \in S_i, s_{i,t} \neq s_{i,j}$  faire
          traiter  $s_{i,t}$ , apparier chaque trame avec  $d_i$  en utilisant  $K_i^{init} \setminus t$ 
        fin pour
      fin pour
    fin pour
  fin pour
```

4.2.3. Méthodes testées

Dans nos expériences, nous avons étudié le comportement des détecteurs et descripteurs SIFT et ORB en utilisant les paramètres suivants :

- pour ORB : $init = 2000$, et $T = \{200, 400, 600, \dots, 2000\}$;
- pour SIFT : $init = 4000$, et $T = \{1000, 2000, 3000, 4000\}$.

Pour chaque méthode, on extrait 1 000 points de chaque trame des séquences d'apprentissage et de test.

4.3. Mesures de Performance

4.3.1. Qualité de segmentation

Pour pouvoir estimer la qualité des résultats obtenus avec chaque variation des méthodes présentées, nous avons utilisé l'index de Jaccard (Everingham *et al.*, 2010) qui peut être interprété comme la mesure de la capacité des méthodes à segmenter correctement les document dans chaque image tout en pénalisant les méthodes qui ne sont pas capables de détecter le document dans certaines images.

Grâce à la vérité terrain, nous utilisons la taille et les coordonnées de l'image de document dans chaque trame f pour commencer par redresser l'image afin que le document soit virtuellement à plat afin de pondérer correctement les surfaces de recouvrement. On obtient alors, pour le quadrilatère solution S_f renvoyé par la méthode

de segmentation, et le quadrilatère cible G_f , deux nouveaux quadrilatères S'_f et G'_f sans effet de perspective. Ceci permet de rendre comparable les mesures entre les différentes trames d'une séquence vidéo. Pour chaque trame f , on calcule l'index de Jaccard (JI) de la façon suivante :

$$JI(f) = \frac{\text{aire}(G'_f \cap S'_f)}{\text{aire}(G'_f \cup S'_f)}$$

où $G'_f \cap S'_f$ est le polygone résultant de l'intersection entre la solution et la vérité terrain, et $G'_f \cup S'_f$ est le polygone résultant de leur union. Le score final de chaque méthode est la moyenne des scores sur l'ensemble de test.

4.3.2. Vitesse de traitement

La vitesse de traitement est estimée en considérant le temps moyen nécessaire à la mise en correspondance de chaque trame de l'ensemble de test avec son modèle. Ceci a été calculé en utilisant un code Python instrumenté, n'utilisant qu'un processeur, sur une machine unique.

4.3.3. Occupation mémoire

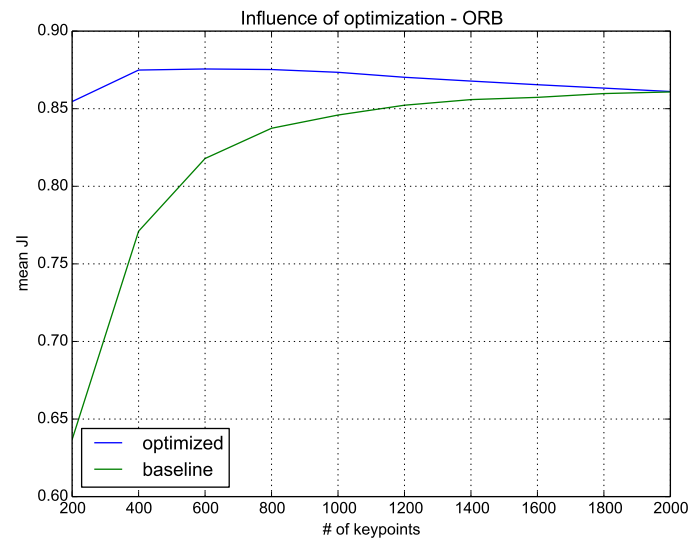
L'impact en mémoire nécessaire peut être directement déduit du nombre de descripteurs conservés pour chaque document modèle.

5. Résultats

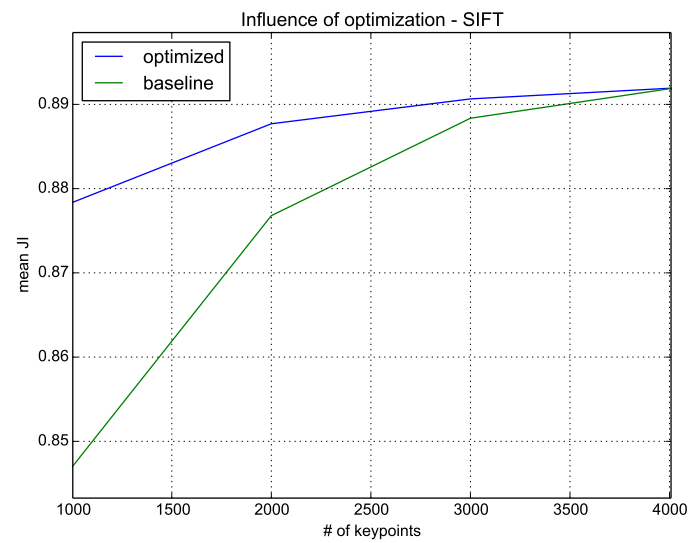
Cette section présente les résultats obtenus avec les descripteurs ORB et SIFT, puis propose des pistes de réflexion pour la bonne mise en œuvre d'un tel mécanisme d'apprentissage.

5.1. ORB

Les résultats montrent un gain important pour ORB lorsqu'on filtre les descripteurs locaux. La figure 4a) montre un gain important lorsque l'on élimine progressivement les points ambigus du modèle, jusqu'à un pic à partir duquel le pouvoir discriminant des descripteurs restants commence à être diminué. Le tableau 2 montre les valeurs obtenues pour l'index de Jaccard. On peut voir qu'un facteur de réduction de 5 à 10 de la taille de l'ensemble des descripteurs peut être atteint avec un impact mineur sur la qualité de la mise en correspondance. Les temps de traitement sont aussi réduits de manière significative, avec un gain de 25% pour des seuils à 200 et 400 point d'intérêts.



a) ORB



b) SIFT

Figure 4. Influence du filtrage sur la qualité de la mise en correspondance.

# descr.	Temps (s)	Jl moyen	
		Référence	Filtrage
2000	0,171	0,861 $\pm$ 0,004	
1800	0,166	0,860 $\pm$ 0,004	0,863 $\pm$ 0,002
1600	0,161	0,857 $\pm$ 0,004	0,865 $\pm$ 0,002
1400	0,155	0,856 $\pm$ 0,004	0,868 $\pm$ 0,002
1200	0,149	0,852 $\pm$ 0,004	0,870 $\pm$ 0,002
1000	0,143	0,846 $\pm$ 0,004	0,873 $\pm$ 0,002
800	0,137	0,837 $\pm$ 0,004	0,875 $\pm$ 0,002
600	0,132	0,818 $\pm$ 0,004	0,876 $\pm$ 0,002
400	0,127	0,771 $\pm$ 0,005	0,875 $\pm$ 0,002
200	0,127	0,637 $\pm$ 0,006	0,855 $\pm$ 0,002

Tableau 2. Variation de la performance en filtrant progressivement 2000 descripteurs locaux avec ORB

# descr.	Temps (s)	Jl moyen	
		Référence	Filtrage
4000	1,781	0,892 $\pm$ 0,004	
3000	1,718	0,888 $\pm$ 0,004	0,891 $\pm$ 0,002
2000	1,624	0,877 $\pm$ 0,004	0,888 $\pm$ 0,002
1000	1,488	0,847 $\pm$ 0,004	0,878 $\pm$ 0,002

Tableau 3. Variation de la performance en filtrant progressivement 4000 descripteurs locaux avec SIFT

5.2. SIFT

À la différence de ORB, SIFT bénéficie moins de l'étape de filtrage, qui freine seulement légèrement la chute de performance comme on peut le voir à la figure 4b), et au tableau 3. Le gain modeste en vitesse ne permet pas de pallier la perte en qualité, ce qui fait de ORB un meilleur candidat dans ce type de situation.

5.3. Influence de l'ensemble d'apprentissage

L'influence de l'ensemble d'apprentissage sur la performance finale du système est significative, comme on peut l'observer à la figure 5. La différence de qualité entre les différents essais (apprentissage sur le fond indiqué, test sur les 4 autres) peut être expliquée par la difficulté très variable des différents fonds de la base de test : lorsqu'on utilise un fond difficile lors de l'apprentissage, ceci rend le test

plus facile dans l'ensemble, et il faut comparer les tendances de courbes plutôt que leurs hauteurs.

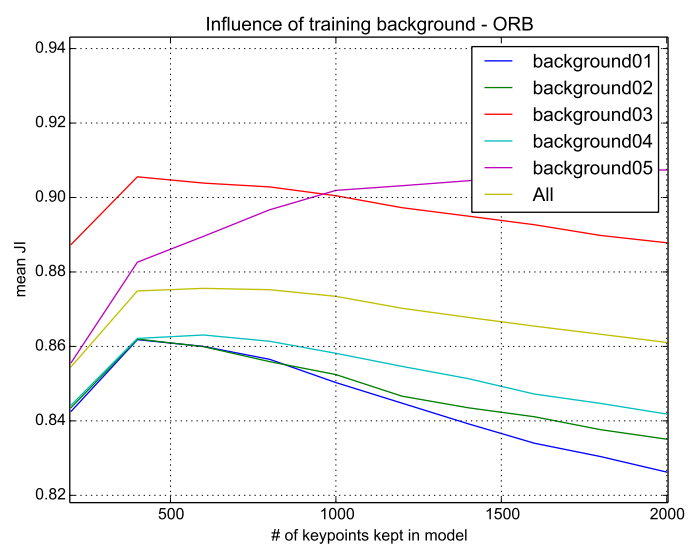
On voit ici que tant pour ORB comme pour SIFT, on devrait éviter les vidéos d'apprentissage qui présentent des objets supplémentaires et des occlusions (fond 05), puisque cela mène à une performance moindre. Dans le cas de ORB, la figure 5a) montre que n'importe quel autre apprentissage apporte des gains importants. Par contre, pour SIFT, il apparaît que les vidéos avec une illumination importante (fond 02) ont un impact négatif fort, tandis qu'il sort renforcé par des fonds avec des illuminations faibles ou qui présentent du flou (fond 03). Des conditions standards (fond 01) semblent également une bonne option pour SIFT.

6. Conclusions

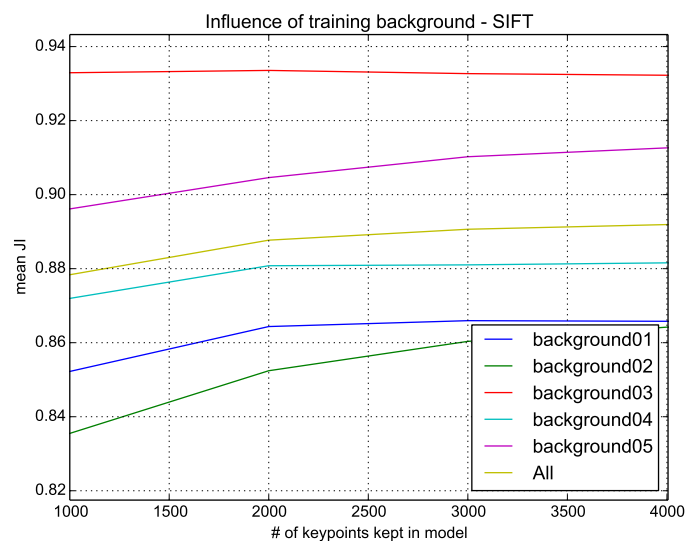
Inspirés par le travail de Kurz et al. (Kurz *et al.*, 2012), nous avons étudié comment une étape de filtrage des points d'intérêt peut influencer la performance d'un système de mise en correspondance d'images de documents. Nous avons évalué cette approche en utilisant la base de données de la compétition SmartDOC d'ICDAR 2015 qui contient près de 25 000 images de documents capturées avec un dispositif mobile, et nous avons montré que le filtrage de points d'intérêt en fonction de la réponse de chacun d'eux sur une base d'apprentissage, apporte des gains de performance importants pour ORB, et plus modestes pour SIFT. Les expériences montrent que si la séquence d'apprentissage est bien choisie, les performances peuvent être encore améliorées. La méthode présentée apporte également des gains significatifs au niveau de la vitesse de traitement des images, et de l'occupation mémoire nécessaire, faisant de cette technique une étape utile dans la préparation d'un processus d'appariement de documents sur plateforme mobile.

Remerciements

This work is partially supported by the People Programme (Marie Curie Actions) of the Seventh Framework Program of the European Union (FP7/2007-2013) under REA grant agreement no. 600388, and by the Agency of Competitiveness for Companies of the Government of Catalonia, ACCIÓ, and the Spanish project TIN2014-52072-P.



a) ORB



b) SIFT

Figure 5. *Influence de l'ensemble d'apprentissage*

7. Bibliographie

- Andoni A., Indyk P., « Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions », *Communications of the ACM – 50th anniversary issue : 1958 – 2008*, vol. 51, n° 1, p. 117-122, January, 2008.
- Antonacopoulos A., Bridson D., Papadopoulos C., Pletschacher S., « A Realistic Dataset for Performance Evaluation of Document Layout Analysis », *Proceedings of the 10th International Conference on Document Analysis and Recognition*, p. 296-300, 2009.
- Bay H., Ess A., Tuytelaars T., Gool L. V., « SURF : Speeded Up Robust Features », *Computer Vision and Image Understanding*, vol. 110, n° 3, p. 346-359, 2008.
- Burie J., Chazalon J., Coustaty M., Eskenazi S., Luqman M., Mehri M., Nayef N., Ogier J., Prum S., nol M. R., « ICDAR2015 Competition on Smartphone Document Capture and OCR (SmartDoc) », *Proceedings of the 13th International Conference on Document Analysis and Recognition*, 2015.
- Calonder M., Lepetit V., Ozuysal M., Trzcinski T., Strecha C., Fua P., « BRIEF : Computing a Local Binary Descriptor Very Fast », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, n° 7, p. 1281-1298, July, 2012.
- Chazalon J., Rusiñol M., Ogier J., Lladós J., « A Semi-Automatic Groundtruthing Tool for Mobile-Captured Document Segmentation », *Proceedings of the 13th International Conference on Document Analysis and Recognition*, 2015a.
- Chazalon J., Rusiñol M., Ogier J.-M., « Improving Document Matching Performance by Local Descriptor Filtering », *Proceedings of the 6th International Workshop on Camera Based Document Image Analysis*, 2015b.
- Dimmick D., Garris M., Wilson C. L., Structured Forms Database, Technical report, National Institute of Standards and Technology, 1991.
- Everingham M., Gool L. V., Williams C., Winn J., Zisserman A., « The PASCAL Visual Object Classes (VOC) Challenge », *IJCV*, vol. 88, n° 2, p. 303-338, 2010.
- Fischler M., Bolles R., « Random sample consensus : a paradigm for model fitting with applications to image analysis and automated cartography », *Communications of the ACM*, vol. 24, n° 6, p. 381-395, June, 1981.
- Ford G., Thoma G., « Ground truth data for document image analysis », *Proceedings of the Symposium on Document Image Understanding and Technology*, p. 199-205, 2003.
- Kurz D., Olszamowski T., Benhimane S., « Representative feature descriptor sets for robust handheld camera localization », *Proceedings of the International Symposium on Mixed and Augmented Reality*, p. 65-70, 2012.
- Lewis D., Agam G., Argamon S., Frieder O., Grossman D., Heard J., « Building a Test Collection for Complex Document Information Processing », *Proc. Int. ACM SIGIR Conf.*, p. 665-666, 2006.
- Lowe D., « Distinctive Image Features from Scale-Invariant Keypoints », *International Journal of Computer Vision*, vol. 60, n° 2, p. 91-110, November, 2004.
- Luqman M., Gomez-Krämer P., Ogier J., « Mobile phone camera-based video scanning of paper documents », *Proceedings of the 5th International Workshop on Camera-Based Document Analysis and Recognition*, p. 77-82, 2013.

- Medvet E., Bartoli A., Davanzo G., « A Probabilistic Approach to Printed Document Understanding », *International Journal of Document Analysis and Recognition*, vol. 14, n° 4, p. 335-347, December, 2011.
- Moraleda J., Hull J., « Toward Massive Scalability in Image Matching », *Proceedings of the International Conference on Pattern Recognition*, p. 3424-3427, 2010.
- Muja M., Lowe D., « Fast approximate nearest neighbors with automatic algorithm configuration », *Proceedings of the International Conference on Computer Vision Theory and Applications*, p. 331-340, 2009.
- Nakai T., Kise K., Iwamura M., « Camera-Based Document Image Retrieval as Voting for Partial Signatures of Projective Invariants », *Proceedings of the 8th International Conference on Document Analysis and Recognition*, p. 379-383, 2005a.
- Nakai T., Kise K., Iwamura M., « Hashing with Local Combinations of Feature Points and Its Application to Camera-Based Document Image Retrieval—Retrieval in 0.14 Second from 10,000 Pages— », *Proceedings of the 1st International Workshop on Camera-Based Document Analysis and Recognition*, p. 87-94, 2005b.
- Rublee E., Rabaud V., Konolige K., Bradski G., « ORB : An efficient alternative to SIFT or SURF », *Proceedings of the International Conference on Computer Vision*, p. 2564-2571, 2011.
- Rusiñol M., Chazalon J., Ogier J., Lladós J., « A Comparative Study of Local Detectors and Descriptors for Mobile Document Classification », *Proceedings of the 13th International Conference on Document Analysis and Recognition*, 2015.