Reconnaissance et extraction de documents. Une application industrielle à la détection de documents semi-structurés

Olivier Augereau, Nicholas Journet, Jean-Philippe Domenger

Reconnaissance et extraction de documents

Une application industrielle à la détection de documents semi-structurés

Olivier Augereau [1], **Nicholas Journet** [2], **Jean-Philippe Domenger** [2]

1. *GESTFORM, 38 Rue François Arago*
   33700 Merignac, France
   oaugereau@gestform.com
2. *Université Bordeaux, LaBRI, UMR 5800*
   F-33400 Talence, France
   journet@labri.fr,domenger@labri.fr

RÉSUMÉ. *Cet article aborde le problème de la reconnaissance d'images de documents semi-structurés. L'objectif est de détecter la présence d'un document dans une image et d'extraire la zone d'intérêt qui le contient. Dans un premier temps, un exemple de document à retrouver est donné en entrée du système et un ensemble de points d'intérêt sont extraits de cette image requête. Ensuite, pour chaque image à comparer, l'ensemble des points d'intérêt sont extraits puis mis en correspondance avec ceux de l'image requête. Cette étape de mise en correspondance permet de calculer la transformation géométrique (translation, rotation, zoom) permettant de localiser précisément l'image requête dans les images à analyser. Deux principales propositions sont faites pour rendre utilisable cette techniques pour la recherche d'image de documents : la sélection de points d'intérêt et l'adaptation de RANSAC.*

ABSTRACT. *This article deals with the problem of recognition of semi-structured documents image. The aim is to detect a document and to extract the region of interest containing it. Initially, an exemple of document is given by the user and a set of interest points are extracted from this query image. In a second step, a set of interest points is extracted from each image to analyse and is matched with the set of the query image. This matching is used to calculate the geometric transformation (translation, rotation, zoom) allowing the registration between the query image and the analysed image. Two main proposals have been made to make this technique usable for documents image matching : the selection of interest points and the adaptation of RANSAC.*

MOTS-CLÉS : *comparaison d'images de documents, points d'intérêt, FLANN, SURF, RANSAC.*

KEYWORDS: *document image comparison, interest points, FLANN, SURF, RANSAC.*

!!!! Journal name undefined !!!! – !!!! Journal language undefined !!!! !!!! Issue number undefined !!!! /!!!! Issue year undefined !!!! , !!!! Issue number undefined !!!!

 !!!! Journal acronym undefined !!!! . !!!! Journal language undefined !!!! !!!! Issue volume undefined !!!! – !!!! Journal language undefined !!!! !!!! Issue number undefined !!!! /!!!! Issue year undefined !!!!

!!!! DOI number undefined (journal DOI type, journal acronym, issue volume, issue number, or article first page or last pasge is missing) !!!!

1. Introduction

Cet article se place dans un contexte industriel. Les entreprises de dématérialisation souhaitent déterminer si des documents semi-structurés tels qu'une carte d'identité, un ticket de train, une facture de téléphone, etc. sont présents sur une page numérisée. Ce que nous appelons "image" dans la suite de l'article est une page numérisée. Une image peut contenir un ou plusieurs documents de différentes natures. Quelques exemples d'images à analyser sont présentés sur la figure 1. Si un document est présent dans une image à analyser, il doit être localisé précisément afin de pouvoir exploiter les informations qu'il contient, telles que le nom, le prénom, etc. Notre problématique se rapproche donc de la recherche de sous-image (*sub-image retrieval*). Les contraintes industrielles imposent de privilégier la précision au rappel : en effet il est préférable que les décisions prises par l'algorithme de localisation soient correctes le plus souvent possible pour ne pas avoir à contrôler la sortie du système. Cet article présente une méthode adaptée à la comparaison de documents semi-structurés. Un document est semi-structuré si l'ensemble des documents de la même "famille" ont une partie de leur structure en commun. Par exemple les cartes d'identité françaises sont des documents "semi-structurés" dans la mesure où une partie de l'information ne change pas d'une carte à l'autre. Seules les informations relatives à la personne (nom, prénom, photo...) changent d'un exemplaire à un autre.

Le processus de comparaison proposé dans cet article repose sur une détection puis une analyse des points d'intérêt sur les différentes images de la base. La nature même des documents à comparer rend complexe cet objectif de comparaison. Par exemple, les documents tels que les pièces d'identité contiennent une photo, des logos, des textures et du texte protégés des contrefaçons. Certaines informations (comme des hologrammes) disparaissent ou sont déformées lors de la numérisation. Des documents tels que les tickets de caisse ou récépissés sont des documents de mauvaise qualité (encre effacée, papier déformé...) et leur numérisation génère une image bruitée. L'étape de numérisation du document physique introduit également un degré de complexité supplémentaire à la recherche et la comparaison de documents. Les documents présents dans l'image peuvent être placés d'une manière quelconque (différentes orientations et positions). Il peut également arriver que l'image du document ne soit pas à l'échelle standard et que certaines parties soient coupées ou recouvertes par d'autres informations. De plus, il peut y avoir des documents de différents types sur une même image.

La problématique abordée dans cet article est de reconnaitre et d'extraire des portions d'images de documents. Les approches classiques de reconnaissance de documents (Chen, Blostein, 2007) reposent davantage sur des techniques de classification supervisées utilisant des caractéristiques basées *a priori* sur le texte, la mise en page ou encore la présence d'illustrations ou de photos. La figure 2 illustre les difficultés par les algorithmes de reconnaissance de caractères à segmenter des images issues de notre base de données.

Une chaine de traitements classique consisterait à segmenter les images puis reconnaitre la classe de chaque document. Récemment, les auteurs de (Rangoni, Belaïd,

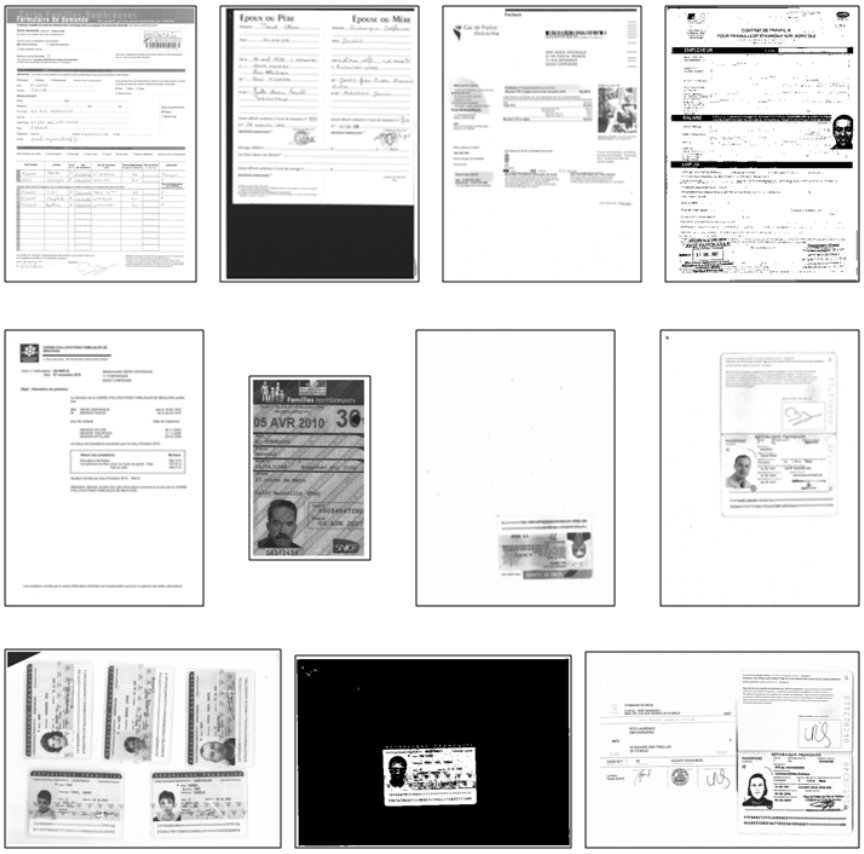


Figure 1. Exemples de documents présents dans la base (de gauche à droite et de haut en bas) : un formulaire, un acte de mariage, une facture, un contrat, une attestation, une carte d'abonnement, un titre de séjour, un passeport français, cinq cartes d'identité françaises, une carte d'identité française et un passeport français accompagné d'un autre document

2012) ont présenté une méthode de segmentation de documents performante basée sur les recouvrements d'espaces blancs. Malgré l'ensemble des prétraitements nécessaires (redressement, suppression des bordures noires, débruitage, etc.) la technique n'est pas robuste aux transformations 2D et mène à une sursegmentation des documents rendant difficile toute étape de classification ultérieure. On peut également remarquer que si les documents présents sur une image ont des orientations différentes, l'étape de redressement échouera.

De manière générale, nous avons observé que les méthodes de segmentation de documents comme celles présentées dans (Chen, Blostein, 2007) ou (Rangoni, Belaïd, 2012) ont tendance à ne pas toujours correctement localiser les zones de texte et que

Figure 2. Résultat d'un algorithme de reconnaissance de caractères illustrant les difficultés rencontrées par ce type de logiciel à segmenter puis reconnaitre le texte sur des documents de type "notes de frais"

l'extraction de la mise en page de documents contenant des tableaux, des tampons, des graphiques, etc. s'opère très difficilement.

L'approche en trois étapes « prétraitement + segmentation + classification » est classique mais n'a jamais été testée sur des bases complexes telles que la notre. La littérature regorge d'articles sur chacune des trois étapes (séparément), pour les rendre génériques et applicable à des bases plus ou moins complexes. Cependant, la combinaison des trois étapes a été proposé uniquement pour des cas particuliers (facture, chèque, etc.). Dans (Augereau *et al.*, 2011), nous avons testé une méthode de classification de ce type. Les limites sont liées à l'apprentissage supervisé volumineux, à la sélection de caractéristiques en fonction des classes de documents à reconnaitre et à la qualité de la segmentation. Dans un contexte industriel, il faut être en mesure de classer plusieurs dizaines de milliers de documents numérisés chaque jour. De ce fait, il paraît impensable de mettre en place une chaine d'analyse ou de traitement nécessitant en permanence de devoir labelliser des données pour une phase d'apprentissage, de paramétrer des prétraitements ou des algorithmes de segmentation.

La possibilité de combiner un système sans prétraitement (segmentation des documents composites, redressement, autres) avec une légère phase d'apprentissage (mode requête par l'exemple) nous a poussés à rechercher la faisabilité d'une approche par recherche de sous-images. Cette approche a été largement abordée dans le domaine des images naturelles grâce à l'utilisation de détecteurs (Tuytelaars, Mikolajczyk, 2008) et de descripteurs (Mikolajczyk, Schmid, 2005) de points d'intérêt. Parmi les applica-

tions, on peut citer les travaux ayant pour objectif de permettre d'aligner des images les unes par rapport aux autres (G. Yang *et al.*, 2007), pour rechercher des logos, (Psyllos *et al.*, 2010) ou encore pour créer des panoramas (Brown, Lowe, 2007).

Les points d'intérêt sont également utilisés sur des images de documents, notamment pour la détection de logos (Zhu, Doermann, 2009), (Rusinol, Lladós, 2009), (Jain, Doermann, 2012), (T. D. Pham, 2003), la localisation de symboles (Nayef, Breuel, 2013), de graphiques (Bhatti, Hanbury, 2011). Le cadre applicatif de ces techniques est la comparaison d'éléments graphiques tels que des logos ou des symboles. Comme l'indiquent les auteurs de (Rusinol, Lladós, 2009), l'un des plus grands défis actuels est de proposer une méthode de reconnaissance sans segmentation de "document complet". Les auteurs de (Takeda *et al.*, 2011) tentent de répondre à cette problématique en proposant une méthode basée sur des points d'intérêt permettant de retrouver un document en temps réel dans une base de données d'un million de documents. L'inconvénient est que cette méthode est entièrement basée sur le texte. Les points doivent être extraits aux centres des mots, il faut donc que le texte soit segmenté et les lettres regroupées efficacement. La présence de bruit, de graphiques, d'images, de texte avec différentes taille de police et la mauvaise qualité des documents impactera directement les résultats de leur méthode qui a été testée uniquement sur des images de documents tels que des articles scientifiques contenant une majorité de texte uniforme. Dans un contexte comme le notre où la base est constituée de documents contenant un grande quantité d'illustrations ou de tableaux cette technique semble difficilement adaptable.

Comme le souligne (Uchiyama, Saito, 2009), la simple transposition de techniques de comparaisons dédiées aux images naturelles basées sur les points d'intérêt tels que SIFT ou SURF donne des résultats peu probants sur les images de document, ce qui explique pourquoi si peu de propositions ont été faites sur la comparaison d'images de documents. Il est donc nécessaire d'adapter ces techniques aux images de documents afin de les rendre utilisables, ce qui est l'objectif de cet article.

La section 2 détaille la chaîne de traitements utilisée de manière standard pour la recherche d'images naturelles par l'exemple (détection de points d'intérêt, mise en correspondance des points et estimation d'une transformation géométrique) telle qu'elle est faite pour la détection d'objets (Lowe, 2004) ou de logo (Jain, Doermann, 2012).

Après avoir montré les limites de la simple transposition de cette chaîne de traitements à la recherche d'images de documents, la troisième section détaille notre méthode, qui est une adaptation des méthodes de détection d'objets ou de logo. Nous verrons donc, comment il est possible d'adapter cette chaîne à notre contexte (reconnaissance de documents complet).

Enfin la quatrième section détaille les résultats obtenus et propose un retour d'expérience sur l'utilisation de notre méthode en production.

2. Reconnaissance d'objet standard

La chaîne de traitements que nous proposons d'utiliser pour la reconnaissance et l'extraction de documents semi-structurés est décrite dans la figure 3. Notre proposition s'inspire de celle faite par (Lowe, 2004). La première étape consiste à sélectionner une image requête. Cette image est simplement un exemple du type d'image que l'on souhaite reconnaitre. Les points d'intérêt des différentes images requêtes et des images à analyser de la base de données sont extraits et décrits à l'aide de la méthode SURF (Bay *et al.*, 2008). Ensuite les points d'intérêt d'une image requête sont mis en correspondance avec les points d'intérêt de chacune des images de la base. Pour cela, l'algorithme de recherche rapide et approximatif de plus proches voisins FLANN (Muja, Lowe, 2009) est utilisé. Enfin, la transformation (modèle à 4 paramètres) permettant de passer de l'image requête à l'image analysée est estimée avec RANSAC (Fischler, Bolles, 1981). Cette opération permet de localiser très précisément le modèle dans l'image requête. Selon Lowe, si au moins 3 mises en correspondance valident la transformation géométrique, l'objet (dans notre cas, le document) est considéré comme étant présent sur l'image. La matrice de transformation géométrique est utilisée pour localiser le document uniquement si ce dernier a été reconnu, *i.e.* si 3 mises en correspondance valident la transformation.

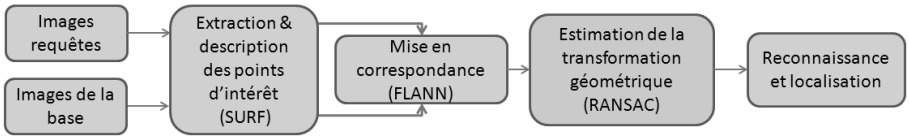


Figure 3. Chaîne de traitement pour la mise en correspondance d'images. Les points d'intérêt de l'image requête et des images de la base à analyser sont extraits et caractérisés avec SURF puis mis en correspondance avec FLANN. La transformation géométrique est ensuite estimée à l'aide de RANSAC

Dans les sous-sections suivantes nous allons présenter chacune des étapes de la chaîne de reconnaissance d'objets standard.

2.1. Détection et description de points d'intérêt

Cette étape consiste à détecter des points d'intérêt dans des images à comparer. Les points sont caractérisés par la configuration des pixels voisins. Le descripteur associé au point doit être le plus robuste possible afin d'être reconnu même si la position et l'orientation du document diffèrent entre les deux images à comparer.

La plupart des détecteurs existants sont invariants en translation, comme par exemple le détecteur de Harris (Harris, Stephens, 1988). Les évolutions telles que Harris-Laplace ou les méthodes basées sur les DoG (Difference of Gaussians) sont quant à elles robustes aux rotations et aux changements d'échelle. Les techniques telles que MSER (Matas *et al.*, 2004) et LLD (Cao, 2008) ont été mises au point dans l'objectif d'être invariantes aux transformations affines. Cependant, SIFT reste la référence en

matière de détection de points d'intérêt. Il combine les DoG qui sont invariants en translation, rotation et mise à l'échelle avec un descripteur basé sur les distributions d'orientations de gradient qui, de plus, est robuste aux changements d'illumination et de points de vues. Depuis, quelques variantes et extensions de SIFT telles que PCA-SIFT (Ke, Sukthankar, 2004), ASIFT (Morel, Yu, 2009) et SURF ont été mises au point. Les différences entre ces techniques sont minimes et portent principalement sur la rapidité d'exécution et leur robustesse au changement d'échelle, à la rotation, au flou gaussien, au changement de luminosité et aux transformations affines. Les auteurs de (Juan, Gwun, 2009) présentent une étude comparative de SIFT, PCA-SIFT et SURF.

L'une des principales qualités du descripteur SURF réside dans sa rapidité de calcul. L'étude comparative (Juan, Gwun, 2009) démontre la supériorité du descripteur SURF par rapport à SIFT et PCA-SIFT en termes de temps d'exécution, de sa robustesse aux changements d'illumination et de son taux de répétabilité. L'algorithme SURF est composé de deux étapes principales. La première consiste à détecter des points d'intérêt sur l'image et la seconde consiste à décrire ces points d'intérêt à l'aide d'un vecteur de 64 caractéristiques.

2.1.1. Détection des points

La détection des points SURF est basée sur la construction d'images intégrales (Viola, Jones, 2001) qui permettent de gagner grandement en temps de calcul. Les zones de fort changement d'intensité des pixels sont recherchées dans l'image. La matrice hessienne, basée sur le calcul des dérivées partielles d'ordre deux, est utilisée pour cela. Si le déterminant de la matrice hessienne est positif, alors les valeurs propres de la matrice sont toutes les deux positives ou toutes les deux négatives, ce qui signifie qu'un extremum est présent. Les points d'intérêt seront donc localisés là où le déterminant de la matrice hessienne est maximal. Le taux de répétabilité de SURF est directement lié au seuillage des maximum locaux de la matrice Hessienne. Des études comparatives montrent que ce taux, par défaut, est déjà bon (Juan, Gwun, 2009) (Kalia *et al.*, 2011). Parce que nous travaillons sur des images de documents noir et blanc, la binarisation des documents permet de rendre le taux de répétabilité moins sensible à ce seuillage car les transitions entre les lettres et le fond sont binaires.

2.1.2. Description des points

Une fois les points d'intérêt extraits, la seconde étape de SURF consiste à calculer le descripteur correspondant. Le descripteur SURF décrit l'intensité des pixels dans un voisinage autour de chaque point d'intérêt. Pour chaque point, la réponse des ondelettes de Haar est calculée dans un voisinage proche. Au final, chacun des points extraits est décrit par un vecteur composé de 64 dimensions.

Les auteurs de (Audithan, Chandrasekaran, 2009) ont démontré que les ondelettes de Haar sont pertinentes pour la description des lettres typographiées dans les images de documents.

2.2. Mise en correspondance des points

Dans cette étape, les points d'intérêt d'une image sont mis en correspondance avec les points d'intérêt d'une autre image afin d'estimer le degré de similitude entre ces deux images. Chaque point d'intérêt de l'image modèle est associé aux deux points d'intérêt de l'image requête qui lui sont les plus proches en termes de distance euclidienne dans l'espace des 64 dimensions (le second plus proche sera utilisé pour l'étape de filtrage que nous détaillons un peu plus loin).

La recherche de PPV peut s'avérer longue si elle est faite de manière exhaustive. L'algorithme FLANN décrit dans (Muja, Lowe, 2009) utilise le principe d'arbres KD aléatoires proposé récemment par (Silpa-Anan, Hartley, 2008) pour faire la recherche approximative de plus proche voisins.

Après la mise en correspondance, chacun des points du modèle est associé aux deux points les plus ressemblants dans l'image requête. On souhaite éliminer le plus possible de fausses mises en correspondance afin de faciliter l'estimation de la transformation. Le premier filtre consiste à supprimer les mises en correspondance dont les 2-ppv dans l'espace de dimension 64 sont trop proches l'un de l'autre. C'est le filtrage par unicité, il permet d'éliminer les mises en correspondance ambiguës. Ensuite les mises en correspondance sont filtrées en fonction de l'échelle et l'orientation. Le rapport de l'échelle et la différence d'orientation des mises en correspondance sont calculés. L'espace des angles est découpé par tranche de 20 degrés et l'espace des échelles par facteurs de $1, 5$. Les mise en correspondances dont l'échelle et la rotation ne concordent pas au vote de l'échelle et l'angle majoritaire, sont éliminées. La figure 5 met en évidence l'intérêt de ces filtrages qui permettent d'éliminer facilement un grand nombre de mauvaises mises en correspondance (passage de l'image a) à l'image b)).

2.3. Estimation de la transformation géométrique

Dans notre cas, les documents sont numérisés à plat, il n'y a pas de distorsion ni de rotation autre que celle dans le plan. Le modèle recherché comporte quatre inconnues : l'angle θ de rotation dans le plan, la translation T_x selon l'axe x, la translation T_y selon l'axe y et la mise à l'échelle α (uniforme en x et y). La matrice de transformation recherchée M_t est de la forme suivante :

$$M_t . \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha.cos(\theta) & -sin(\theta) & T_x \\ sin(\theta) & \alpha.cos(\theta) & T_y \\ 0 & 0 & 1 \end{bmatrix} . \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix}$$

La difficulté de cette partie est de trouver une transformation géométrique parmi les différentes mises en correspondance alors qu'il reste des mises en correspondance erronées et que le document recherché peut ne pas être présent sur l'image.

Pour ces deux raisons, la méthode des moindres carrés ne peut pas être utilisée car elle serait perturbée par les mises en correspondance erronées (valeurs aberrantes) et

trouverait une transformation même quand le modèle n'est pas présent. L'estimation de la transformation géométrique M_t doit donc être faite à l'aide d'un algorithme capable d'estimer un modèle sans prendre en compte les valeurs aberrantes (aussi appelées *outliers*). C'est la raison pour laquelle l'algorithme RANSAC est très souvent utilisé pour l'estimation de transformation géométrique (Sur *et al.*, 2011).

L'algorithme RANSAC s'articule en deux étapes principales. 1) Le sous-ensemble le plus petit possible permettant d'estimer le modèle géométrique est sélectionné aléatoirement. 2) On cherche d'autres éléments validant le modèle. Ces éléments sont appelés *inliers*. S'il n'y a pas suffisamment d'*inliers*, l'algorithme retourne en 1), sinon le modèle est validé. Si aucun modèle n'est validé après MAX_ITER essais, l'algorithme s'arrête. MAX_ITER est fixé empiriquement à 200.

En ce qui concerne les paramètres, si MAX_ITER est trop bas le meilleur ensemble de mises en correspondance peut être manqué. Mais si sa valeur est trop grande, le temps d'exécution sera allongé. Un autre seuil est fixé dans RANSAC : $DIST\_VALID$. C'est la distance maximale entre la position attendue de la mise en correspondance donnée par la transformation et la position réelle de la mise en correspondance. Cette valeur représente la tolérance du processus de mise en correspondance. Puisque la transformation géométrique est supposée être linéaire (on considère qu'il peut y avoir uniquement une translation, une rotation dans le plan et un changement d'échelle) la valeur de ce seuil ne doit pas être trop haute. Cependant, nous travaillons avec des images qui sont relativement larges, les documents A4 sont des documents numérisés à 200 dpi. Nous avons expérimentalement déterminé $DIST\_VALID$ à 10 pixels. Cette valeur permet d'être tolérant à de légères déformations. Si cette valeur est trop élevée, cela risque d'augmenter le nombre de mauvaises mises en correspondance. La figure 4 illustre un exemple d'utilisation de RANSAC.

La figure 5 montre l'impact de RANSAC sur les mises en correspondance (passage de l'image b) à l'image c)).

Enfin, si au moins t mises en correspondances sont validées par le modèle géométrique, le document est considéré comme étant présent dans l'image. Nous avons choisi $t = 3$, selon les recommandations de (Lowe, 2004).

2.4. Performance de la méthode standard appliquée à des images de documents semi-structurés

La technique standard a été appliquée à une base de données d'images de documents réels. Nous appellerons cette première base BD_NB. Les images de cette base ont été numérisées et binarisées à la volée par les scanners afin d'être stockées dans des fichiers de format TIF groupe 4. Dans cette base, la plupart des images sont au format A4 et toutes ont une résolution de 200 dpi. Chaque image ne contient qu'un seul document. Cette base contient 2 155 images de documents. Parmi ces documents nous nous intéresserons à 7 types de documents en particulier. Ces documents sont les suivants : 483 cartes d'identité françaises, 89 passeports français, 35 tickets de train

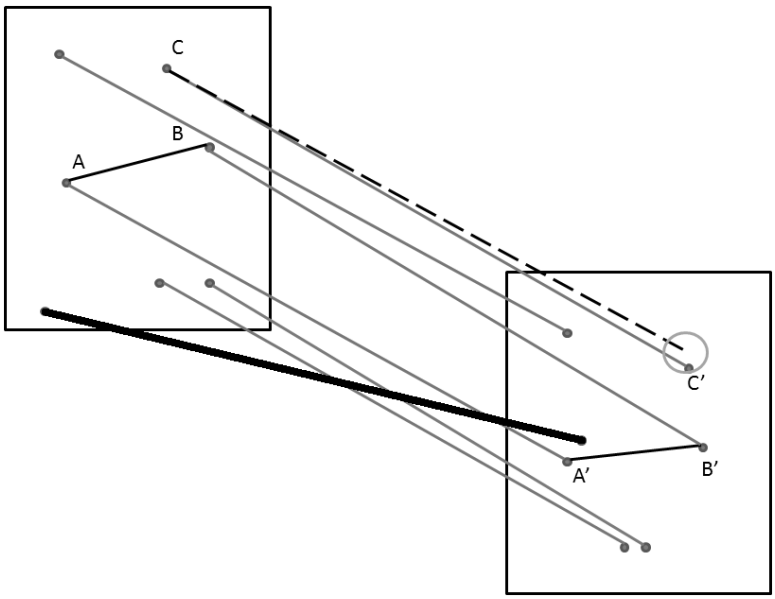


Figure 4. Utilisation de RANSAC. La matrice de transformation M_t est celle qui transforme le vecteur AB en $A'B'$. Si le point C' est à une distance inférieure à $DIST\_VALID$ de la transformation de C pat M_t, alors le point est validé et fait partie des inliers. Cette distance est symbolisée par le cercle sur la figure. Sinon il n'est pas validé et fait partie des outliers (symbolisé par un trait noir épais)

S.N.C.F., 228 bordereaux d'envoi, 58 tickets de restaurant "Challenger", 58 factures "Orange" et 41 reçus "American Express". On peut noter que certaines classes possèdent des similarités visuelles entre elles. Par exemple, les bandes MRZ et les photos sont fortement similaires entre une pièce d'identité et un passeport. De même, les factures orange et american express possèdent des zones similaires : lignes de séparation, entêtes, tableaux, chiffres en bas de la page...

Les 1 257 images de documents restantes correspondent à 281 types de documents. La figure 7 illustre la présence d'images confusives parmi les 1 257 documents faisant parti de la classe de rejet. Par exemple, cette classe de rejet contient des tickets de restaurant se trouvant être très similaires à ceux de la classe "Chalenger". Elle contient également des images de factures, des tickets de transports similaires aux autres classes recherchées.

Le nombre moyen de points d'intérêt est de 7 098 pour les cartes d'identité, 16 158 pour les passeports et 13 391 pour les autres documents. La carte d'identité et le passeport utilisés comme modèles possèdent respectivement 7 687 et 10 059 points d'intérêt.

Pour une classe donnée, l'objectif est de retrouver l'ensemble des documents dans la base en donnant un exemple d'image. Cette image requête, préalablement choisie

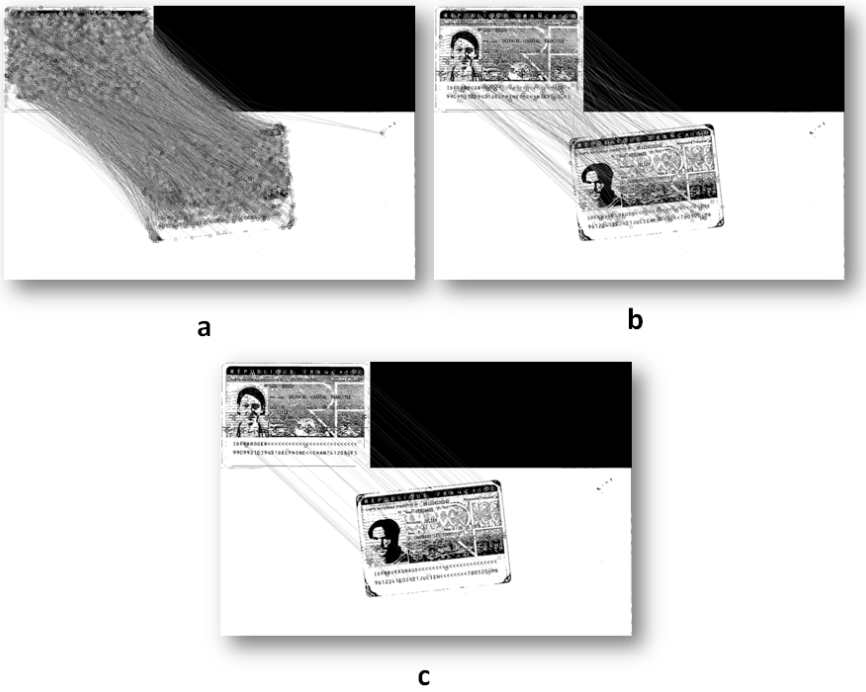


Figure 5. Mise en correspondance de documents. a) Il y a 7 687 *mises en correspondance. b) Après filtrage, il y a* 235 *mises en correspondance. c) Après RANSAC il ne reste plus que* 40 *mises en correspondance. La transformation trouvée a les paramètres suivants : rotation* = 5.665 *degrés, mise à l'échelle* = 1.003*, translation horizontale* = 801.7 *pixels et translation verticale* = 108.3

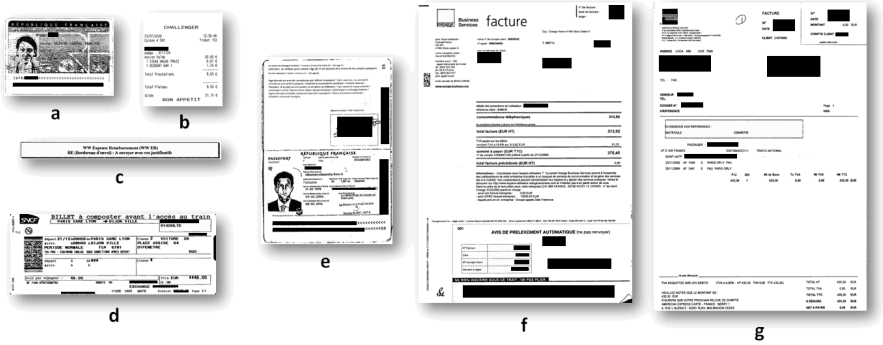


Figure 6. Images utilisées comme requête pour la détection des 7 types de documents. a) Carte d'identité "ID", b) ticket de restaurant "Challenger", c) un résumé de note de frais "Bordereau", d) un ticket de train "SNCF", e) un passeport français, f) une facture "Orange" et g) un reçu de paiement "American". Les zones noires ont été ajoutées pour des raisons de confidentialité

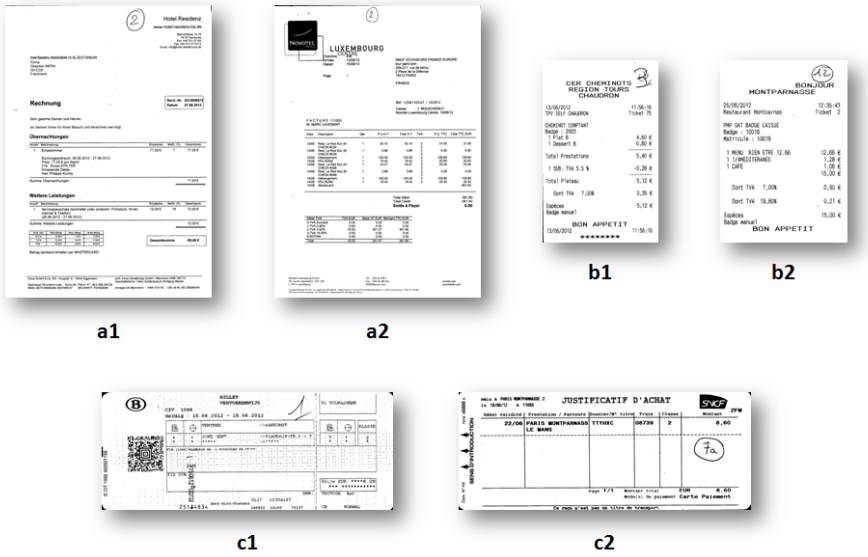


Figure 7. Exemples d'images appartenant à la classe de rejet composée de 1257 documents. On remarque que les documents a1, a2 sont similaires aux factures orange et american express de la figure 6. Les images b1 et b2 sont similaires aux tickets de restaurants challenger de la figure 6. Enfin, les tickets c1 et c2 ont une mise en page ou un logo similaire aux documents de la classe SNCF

par l'utilisateur, est comparée à chaque autre image de la base. Si on arrive à mettre en correspondance une image de la base avec l'image requête alors le document est détecté comme étant présent puis est extrait. Il est donc naturellement préférable de choisir une image requête peu bruitée. Les documents utilisés comme requête sont présentés sur la figure 6.

Tous les tests de ce chapitre ont été effectués sur un ordinateur Intel Core 2 Duo, 2 GHz. Le tableau 1 présente les résultats de la recherche de 7 types d'images dans la base BD_NB. La recherche de chaque classe est indépendante des autres. Ce tableau montre clairement les limites d'une simple transposition d'une méthode de reconnaissance d'objets à des images de documents. Les performances en termes de précision sont extrêmement mauvaises, une moyenne de 17,0 % est obtenue. La grande majorité des documents détectés sont des "faux positifs", c'est-à-dire des documents détectés à tord. Ceci s'explique principalement à cause de la grande quantité de points extraits sur chaque image, ce qui complexifie la mise en correspondance des points. Les mauvaises mises en correspondance sont tellement nombreuses que le système arrive à former un consensus de 3 mises en correspondance sur la plupart des images. On peut également, grâce à ce tableau, établir un lien entre le nombre de points des images requêtes et la précision de la mise en correspondance.

Tableau 1. Rappel, précision et moyenne du temps d'exécution par image en utilisant la technique standard sur la base d'images de documents BD_NB pour la détection de 7 types de documents. De très nombreuses mauvaises détections sont faites, la précision est extrêmement basse

Type	Nb images	Nb points	Rappel	Précision	Tps / image
Cartes Id	483	7687	0,99	0,42	2,9 s
Passeport	89	10059	1,00	0,12	3,9 s
SNCF	35	4934	1,00	0,05	2,4 s
Bordereau	229	751	0,99	0,46	1,3 s
Challenger	58	1573	1,00	0,14	1,6 s
Orange	58	14741	1,00	0,04	5,0 s
American	41	8714	0,95	0,03	2,8 s

Les images de documents contiennent de nombreux points d'intérêt principalement à cause du texte contenu dans les documents. De plus, les images sont numérisées à 200 dpi, un document A4 mesure environ 1 654 pixels de largeur et 2 339 pixels de hauteur. Cette grande quantité d'informations rend l'opération de mise en correspondance des points complexe car le risque de faire de mauvaises mises en correspondance est plus élevé.

La figure 8 présente les 2 cartes d'identité qui n'ont pas été retrouvées en utilisant cette approche. Comme énoncé précédemment, nous avons une contrainte forte visant à obtenir une précision la plus proche possible de 100 %.

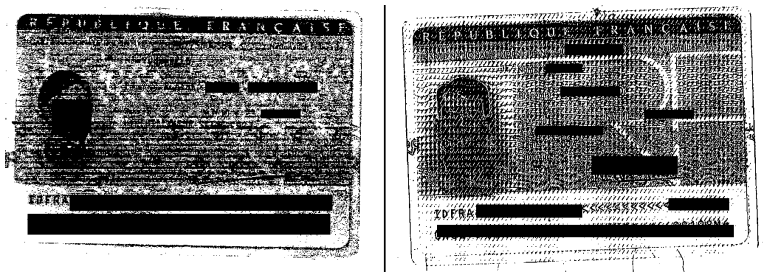


Figure 8. Exemples d'images non détectées. Ces deux images ont un bruit très marqué et ne sont pas détectées

Comme le soulignent les auteurs de (Uchiyama, Saito, 2009), les techniques basées sur les points d'intérêt tels que SIFT ou SURF ne sont pas directement transposables aux images de documents car des patterns répétitifs (comme les lettres des mots) apparaissent fréquemment dans les images de documents. Concrètement, ces patterns sont des petites portions d'images qui se répètent. Si une image présente de nombreux patterns répétitifs, cela peut engendrer un problème d'autosimilarité, car ces portions d'images qui se répètent risquent d'être confondues les unes avec les autres. De plus, de nombreux points sont extraits sur les images de documents, ce qui engendre un très grand nombre de mauvaises mises en correspondance. Le problème de l'application

de la technique classique exposée précédemment vient de la complexité de la mise en correspondance et de la validation d'une transformation parmi un ensemble très vaste. Si aucune précaution n'est prise, de nombreuses images seront alors détectées comme étant identiques à l'image requête.

Nous allons proposer dans la section suivante plusieurs solutions permettant d'améliorer à la fois la précision de la mise en correspondance d'images de documents mais aussi le temps d'exécution.

3. Reconnaissance d'objets adaptée aux images de documents

Dans cette section, nous présentons des améliorations pour adapter la technique de reconnaissance d'objets aux images de documents. Dans un premier temps, nous nous sommes consacrés à l'amélioration de la précision de la technique standard. En effet, comme nous l'avons vu dans la partie précédente de nombreuses images sont détectées à tord (faux positifs). Pour remédier à cela, nous proposons une étape supplémentaire permettant de sélectionner des points d'intérêt pertinents de l'image requête et ainsi de diminuer le temps d'exécution mais également de diminuer la probabilité de faire de mauvaises mises en correspondance. Nous proposons également une version adaptée de RANSAC dans la sous-section 3.2. Nous montrons également quel est l'impact de t (le nombre minimal de mises en correspondance pour valider une transformation) sur le rappel et la précision du système.

La nouvelle chaîne de traitements est présentée sur la figure 9.

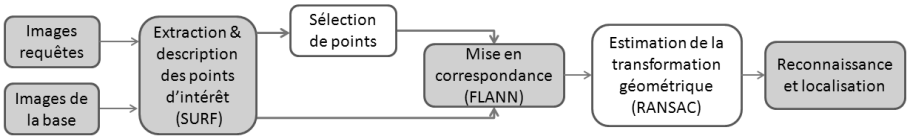


Figure 9. Chaîne de traitement adaptée pour la mise en correspondance d'images de documents. Toutes les étapes sont les mêmes que la méthode standard sauf celles sur fond blanc. Une étape supplémentaire permet de sélectionner certains des points d'intérêt de l'image requête. La transformation géométrique est estimée à l'aide d'une version adaptée de RANSAC qui permet d'augmenter les contraintes géométriques

3.1. Sélection automatique de points d'intérêt

Nous proposons une amélioration visant à diminuer le temps d'exécution et à augmenter la précision du système. L'objectif est de limiter le nombre de points d'intérêt utilisés pour caractériser les images requêtes. En effet, grâce aux tests effectués dans la sous-section 2.4, nous avons vu que l'augmentation du nombre de points d'intérêt décrivant les requêtes a tendance à faire augmenter le temps d'exécution et à faire diminuer la précision. Ceci s'explique par le fait que l'augmentation du nombre de

points d'intérêt induit une augmentation des possibles mises en correspondance et donc l'augmentation du risque de faire de mauvaises mises en correspondance.

Dans une première version de ces travaux (Augereau *et al.*, 2012), nous avons proposé une solution consistant à tracer un rectangle blanc sur les régions d'images requêtes afin de n'extraire les points d'intérêt que sur les zones les plus pertinentes de l'image requête. L'inconvénient est qu'il peut s'avérer complexe de définir manuellement de telles zones et particulièrement pour un opérateur qui n'est pas spécialiste en analyse d'images. Nous proposons ici une évolution permettant d'éviter cette sélection.

La solution que nous proposons consiste à fournir 5 exemples d'images au lieu d'une. Les points d'intérêt de l'une des 5 images requêtes sont mis en correspondance avec les 4 autres images en utilisant RANSAC adapté. Si une transformation est validée, l'ensemble des points d'intérêt mis en correspondance est alors sélectionné. En itérant ainsi sur les 4 nouvelles images, on conserve finalement l'union des points d'intérêt utilisés pour l'ensemble des mises en correspondance faites. Ceci permet donc de garder uniquement les points d'intérêt utiles à la détection de documents du même type.

La figure 10 montre l'impact de la sélection de points d'intérêt sur 5 modèles. La sélection permet en moyenne de diviser le nombre de points d'intérêt par image requête d'un facteur 4,11.

3.2. Adaptation de RANSAC pour une reconnaissance précise des images de documents

Plusieurs adaptations de RANSAC ont déjà été proposées dans la littérature. Les auteurs de LO-RANSAC (Chum *et al.*, 2003) et (Raguram *et al.*, 2008) proposent des adaptations permettant d'accélérer l'exécution de RANSAC. Pour faire la mise en correspondance de points, les auteurs de SCRAMSAC (Sattler *et al.*, 2009) proposent une adaptation utilisant les points du voisinage pour contrôler leur bonne mise en correspondance. SCRAMSAC permet de renforcer la décision d'une mise en correspondance entre points en étudiant le contexte local.
Ces variantes de RANSAC pourraient être utilisées pour améliorer notre adaptation en optimisant les temps de calcul de notre méthode ou pour renforcer les décisions de mise en correspondance. Les deux adaptations que nous avons faites sont indispensables : il ne faut pas considérer des points trop proches au sein d'une image au risque de mal estimer la transformation géométrique. Il faut également supprimer les mises en correspondances qui convergent vers des points trop proche car cela mène généralement à une mauvaise reconnaissance.

L'utilisation de telles méthodes sur des images de documents nécessite une adaptation car les spécificités de nos images (répétition de lettres, motifs texturés...) nécessitent de limiter les correspondances locales. Les tests réalisés dans la section précé-

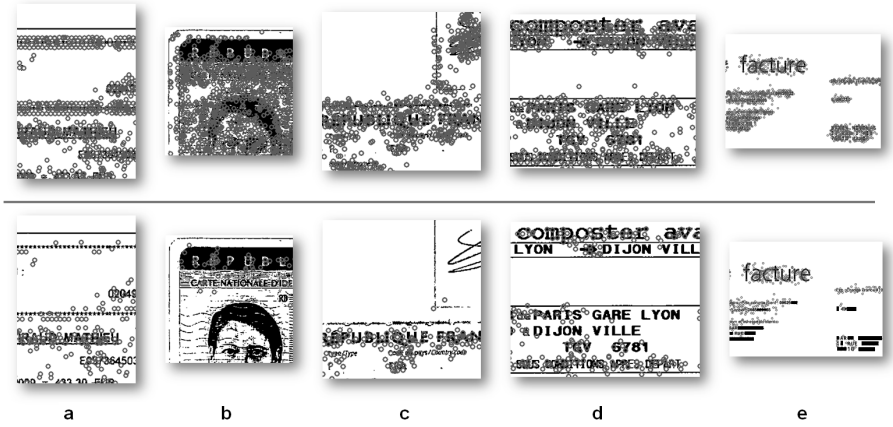


Figure 10. Sélection de points d'intérêt. Les cercles dessinés sur chaque image correspondent à des points d'intérêt. Sur la ligne du dessus figure les images requêtes avant la sélection de points. De nombreux points sont présents. En dessous, sont présentées les mêmes images requêtes après sélection des points d'intérêt. Le nombre de points a largement diminué. On observe que les points d'intérêt situés sur des parties variables de l'image ne sont pas sélectionnés. On peut citer par exemple : b) la photographie de la carte d'identité, c) la signature du passeport, d) le nom des villes sur un ticket de train et d) l'adresse de destination sur une facture. Les nombres de points d'intérêt avant et après sélection sont détaillés dans le tableau 2

dente montrent qu'il y a un nombre important de fausses détections qui se font à cause des mises en correspondances locales.

Nous proposons donc une adaptation de cette famille de méthodes afin de relativiser l'importance du contexte local. Nous illustrerons cette adaptation en modifiant l'algorithme original de RANSAC. Ceci se traduit par l'introduction de seuils interdisant la mise en correspondance de plusieurs points dans un espace trop réduit de l'image.

Afin d'augmenter la précision de RANSAC nous avons ajouté des contraintes géométriques au sein de l'algorithme. Notre version adaptée de RANSAC est détaillée dans l'algorithme 1. Les entrées de l'algorithme sont S_Q et S_D qui représentent l'ensemble des points d'intérêt de l'image requête S_Q mis en correspondances avec les points de l'image de la base à analyser S_D. En sortie de l'algorithme on obtient I, un ensemble d'*inliers*, c'est-à-dire un ensemble de mises en correspondance validant une même transformation géométrique M.

Dans cette version adaptée apparaissent deux nouveaux seuils : MIN_NORME et $MIN\_DVALIDE$. On utilise MIN_NORME pour ne pas calculer la matrice de transformation à partir des points qui sont trop proches car cela risquerait d'induire un calcul approximatif (par analogie, il est préférable d'estimer le coefficient directeur

Algorithm 1 Adapted RANSAC

INPUT S_Q, S_D
OUTPUT I, M
$S_I \leftarrow \emptyset, iter \leftarrow 0$
while $iter < MAX\_ITER$ **do**
  Soient P_{Q1} & P_{Q2} 2 points aléatoirement choisis dans S_Q.
  Soit P_{D1} & P_{D2} la mise en correspondance de P_{Q1} & P_{Q2} dans S_D
  if $(||\overrightarrow{P_{Q1}P_{Q2}}|| < MIN\_NORM)||(||\overrightarrow{P_{D1}P_{D2}}|| < MIN\_NORM)$ **then**
    $iter \leftarrow iter + 1$
    continue
  Soit M_t la matrice de transformation courante
  Soit I_t l'ensemble d'*inliers* courant
  $M_t \leftarrow transfo(\overrightarrow{P_{Q1}P_{Q2}}, \overrightarrow{P_{D1}P_{D2}})$
  $I_t \leftarrow \emptyset$
  for each $P_{Qi} \in S_Q \backslash \{P_{Q1}, P_{Q2}\}$ **do**
    Soit P_{Di} la mise en correspondance de P_{Qi} dans S_D
    $I_t \leftarrow$ TESTAJOUTINLIER$(I_t, P_{Qj}, P_{Qi}, P_{Dj}, P_{Di})$
  if $Card(I_t) > Card(I)$ **then**
    $I \leftarrow I_t, M \leftarrow M_t$
  $iter \leftarrow iter + 1$
return I, M
function TESTAJOUTINLIER$(I_t, P_{Qj}, P_{Qi}, P_{Dj}, P_{Di})$
  for each $I_j \in I_t$ **do**
    if $(||\overrightarrow{P_{Qj}P_{Qi}}|| < MIN\_DVALID)||(||\overrightarrow{P_{Dj}P_{Di}}|| < MIN\_DVALID)$ **then**
      return I_t
  $P'_{Qi} \leftarrow M_t[P_{Qi}]$
  if $||\overrightarrow{P_{Di}P'_{Qi}}|| < MAX\_DIST$ **then**
    $I_t \leftarrow I_t \cup \overrightarrow{P_{Di}P'_{Qi}}$
  return I_t

d'une droite en prenant deux points éloignés). Le second seuil $MIN\_DVALIDE$ est utilisé pour ne pas valider des *inliers* qui sont trop proches. Ces deux seuils permettent d'éviter des mauvaises validations comme celles de la figure 12. MIN_NORME et $MIN\_DVALIDE$ ont été fixés à 5 pixels. Ces seuils ont été validés par des tests effectués sur des documents de différentes tailles. La figure 11 illustre l'utilisation des deux nouveaux seuils.

Enfin, pour prendre encore plus en compte la contrainte géométrique entre points d'intérêt, nous avons choisi d'augmenter le nombre minimum d'inlier t pour valider la transformation. Lowe le fixait à 3 dans les images naturelles. Nous avons choisi de le fixer à 8 pour les images de document. On rappelle qu'après l'application de RANSAC, la transformation n'est validée que si : $Card(I)$ >= t. RANSAC adapté permet de grandement augmenter la précision grâce aux contraintes géométriques supplémentaires et à légèrement diminuer le temps d'exécution.

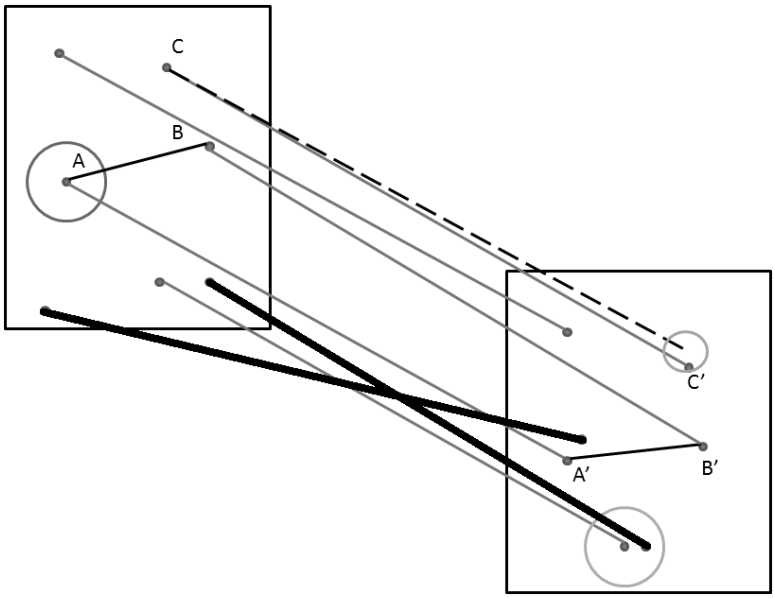


Figure 11. Utilisation de RANSAC adapté. Le fonctionnement global est le même que celui de la figure 4. La norme minimale pour calculer la matrice de transformation MIN_NORME est symbolisée par le petit cercle. La distance minimale pour pouvoir valider un nouveau point $MIN\_DVALIDE$ est symbolisée par les grands cercles

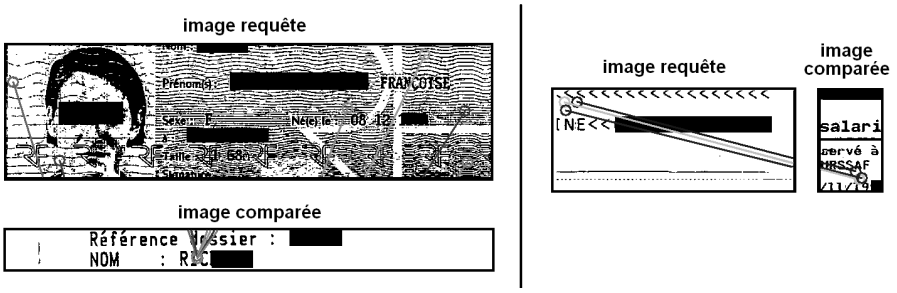


Figure 12. Exemple de mauvaises détections évitées grâce à la version de RANSAC adaptée. a) 6 points éloignés sont mis en correspondance avec deux points proche. b) 3 points proches sont mis en correspondance avec 3 autres points proches. Dans les deux cas il y a trop de mises en correspondance faites dans un espace trop réduit de l'image. RANSAC adapté interdit de telles mises en correspondance et réduit ainsi le nombre de mauvaises mises en correspondance

4. Tests sur bases réelles avec la méthode adaptée

4.1. Application à une base de documents noir et blanc

Cette chaîne de traitements adaptées aux spécificités des images de documents a été testée sur la même base d'images que dans la partie précédente (BD_NB) et selon le même protocole. Le tableau 2 présente les résultats obtenus. On remarque qu'il y a une très légère diminution du rappel mais en contrepartie, une forte hausse de la précision est réalisée. Dans le test de détection standard, le rappel est fort car un très grand nombre d'images sont détectées, à tort, comme similaire à l'image requête. En moyenne, la précision est multipliée par un facteur 5,7 et le temps d'exécution est divisé par 2,0. Les gains maximaux sont obtenus pour les passeports et les factures "Orange" qui sont les documents ayant le plus de points d'intérêt et qui étaient les plus mal détectés par la méthode d'origine.

Tableau 2. Performances de détection de 7 types de documents en utilisant la technique avec RANSAC adapté sur la base BD_NB. On remarque une large diminution du nombre de mauvaises détections

Type	Nb im	Nb pts	Nb pts sélec	Rappel	Précision	Tps / im
Cartes Id	483	7687	306	0,92	1,00	1,01 s
Passeport	89	10059	1284	0,98	1,00	1,14 s
SNCF	35	4934	1846	0,97	1,00	1,58 s
Bordereau	229	751	281	0,90	1,00	1,06 s
Challenger	58	1573	626	1,00	0,88	1,15 s
Orange	58	14741	4444	1,00	0,95	2,30 s
American	41	8714	3017	0,85	0,97	1,49 s
Moyenne				0,93	0,99	

La précision est très bonne mais n'est pas totalement parfaite pour certaines classes d'images. Des mauvaises détections apparaissent là où des documents de classes différentes présentent des similarités fortes. La figure 13 montre des exemples de mauvaises détections. La classe "Challenger" contient des tickets de restaurants, les 6 mauvaises détections se produisent avec d'autres tickets de restaurant qui ont une mise en page et quelques mots clés similaires. Les factures "Orange" sont confondues avec 3 tickets "Air France" parce que le même mot (écrit avec une grande police) est présent sur les deux images. Un autre exemple de mauvaises détections est observé avec le logo des reçus "American" qui se retrouve être identique avec celui utilisé sur des factures "American Express". Ceci montre les limites de notre méthode dans le cas de figure où certaines images de la classe de rejet peuvent être confondues avec les images des classes recherchées. Dans la section suivante, nous détaillons des tests effectués en production montrant qu'il est possible, dans une certaine mesure, d'adapter le seuil lié aux *inliers* pour diminuer les confusions.

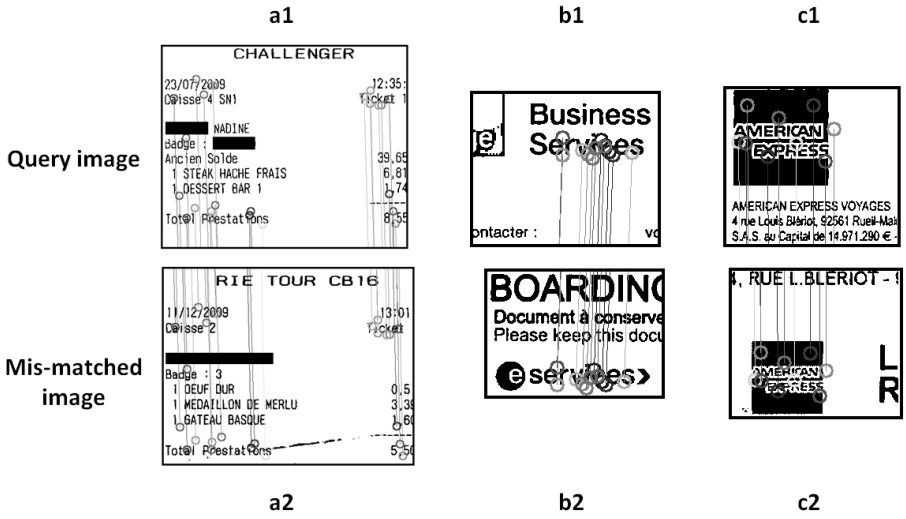


Figure 13. Exemples de mauvaises détections. Le document requête"Challenger" a1) est mis en correspondance avec l'autre ticket de restaurant a2). La facture "Orange" b1) est mise en correspondance avec le ticket "Air France" b2). Le reçu "American" est mis en correspondance avec une facture "American express" c2)

4.2. Retour sur l'utilisation de cette méthode en production

La méthode détaillée dans cet article a été testée en production dans une entreprise de dématérialisation où les documents physiques sont numérisés pour ensuite être indexés. Cette mise en production nous a permis d'obtenir un retour sur la qualité de notre proposition et plus précisément du lien entre la définition des paramètres et les objectifs souhaités en termes de précision et rappel. En effet, l'amélioration du rappel ou de la précision est possible en diminuant ou augmentant certains des seuils de notre proposition.

Les tests en production ont montré que c'est principalement une précision à 100 % qui était espérée (quitte à faire baisser le rappel). Pour cela les opérateurs ayant manipulé notre système de comparaison d'images de documents ont montré que ce sont les seuils liés aux contraintes géométriques globales (lors de la comparaison des points d'intérêt) qui devaient être ajustés. Plus exactement, ils ont eu à modifier les seuils t (nombre d'*inliers* validant la transformation) et $MIN\_DVALIDE$ interdisant la mise en correspondance de points d'intérêt trop près les uns des autres. En augmentant ces deux seuils (de manière empirique), des précisions proches de 100 % ont été obtenues. De fait, le rappel a systématiquement diminué, ceci notamment parce que les documents bruités ne sont plus détectés. La figure 14 montre des exemples de bonnes détections obtenues en production sur différents documents.

Afin de montrer l'impact du seuil t (si plus de t *inliers* sont trouvés, la transformation est validée), le rappel et la précision ont été calculés pour trois types d'images en

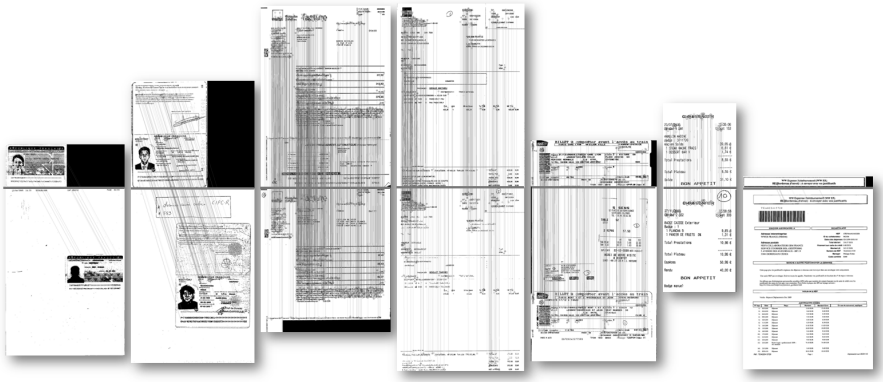


Figure 14. Exemples de bonnes détections pour chacune des 7 images requêtes

faisant varier t. On observe sur la figure 15, les courbes correspondant aux résultats. On en remarque sur ces courbes ce qui a été observé empiriquement par les opérateurs en production. Une valeur de t faible induira une précision faible et un rappel élevé tandis qu'une valeur de t élevée impliquera un rappel faible et une précision élevée. Cependant le rappel chute beaucoup moins rapidement que la précision n'augmente. Nous avons également fait apparaître sur la figure 15, pour $t = 8$ (valeur que nous préconisons), les rappels et précisions obtenus. On observe qu'à chaque fois la précision est de 100 % et que le rappel est encore relativement bon.

Nous avons également observé que dans le cas de figure où un opérateur souhaite retrouver une classe de documents trop similaire visuellement à une autre classe de la base il était possible, en augmentant le seuil t, de diminuer le nombre de confusions.

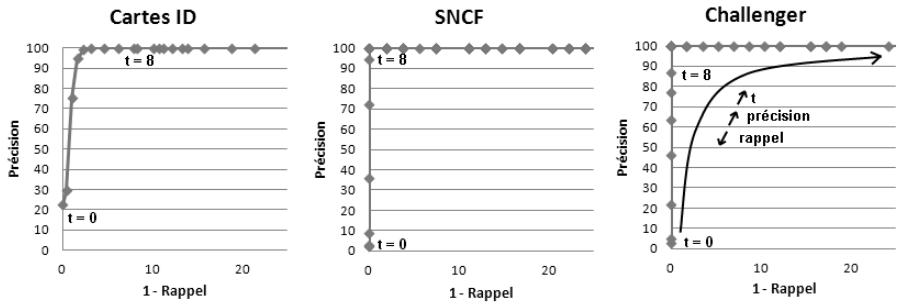


Figure 15. Impact de t sur le rappel et la précision. La précision augmente avec t tandis que le rappel diminue. Sur le graphique, plusieurs points sont superposés et l'abscisse ne représente pas le rappel mais (1-rappel)

4.3. Gestion de la multidétection

La plupart des documents sont composés d'une ou plusieurs pages et donc d'une ou plusieurs images. Cependant il arrive également qu'une seule image puisse contenir plusieurs documents. Le fait que plusieurs documents d'un même type soient présents sur une même image peut complexifier la mise en correspondance. En effet puisque le même type de document est présent plusieurs fois sur une même image, la mise en correspondance des points d'intérêt de l'image requête risque d'être dispersée entre les différents exemplaires du document. Cependant puisque le nombre de points d'intérêt est grand, la mise en correspondance reste possible. La méthodologie pour faire de la multidétection est la suivante. On essaye de mettre en correspondance l'image requête avec les images de la base. Si le document est détecté, il est localisé grâce à la matrice de transformation et la zone correspondante au document est supprimée de l'image (remplacée par un rectangle blanc). Ensuite, l'image est traitée de nouveau jusqu'à ce qu'aucun document ne soit détecté. La figure 16 illustre ce principe.

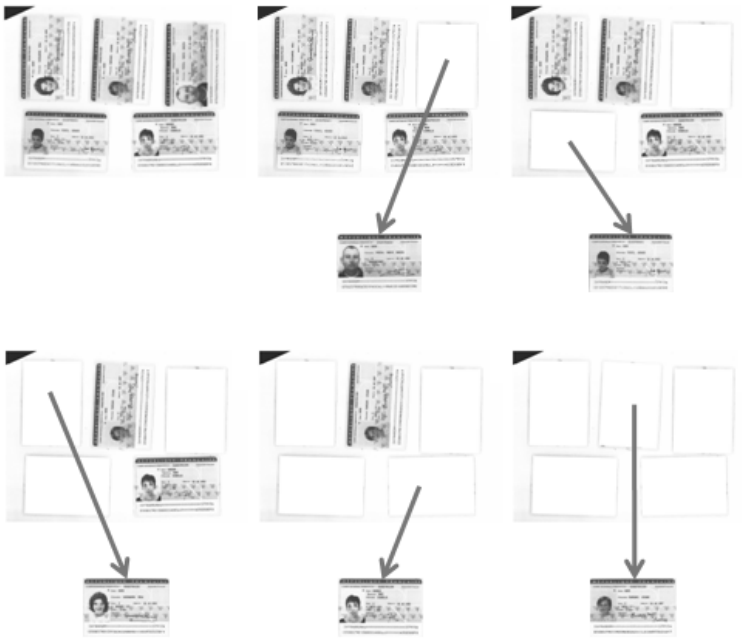


Figure 16. Multidétection. On cherche un modèle dans chaque image. Si un modèle est trouvé, il est supprimé de l'image puis on cherche à nouveau si un modèle est présent sur l'image. On réitère jusqu'à qu'on ne trouve plus de correspondance

Nous avons testé ce principe sur une base où chaque image contient plusieurs documents de même type. La nature des images de la base BD_Multi est la même que celle de la base BD_NdG. La base est composée d'un total de 32 images contenant

89 pièces d'identité. Le tableau 3 montre que la multidétection fonctionne très bien grâce à notre méthodologie.

Tableau 3. Les documents multiples dans les images sont trouvés par l'algorithme. Il peut tout de même manquer des documents pour les mêmes raisons que précédemment, le document manquant est un document très bruité

Modèles	Nb d'im	Nb de doc	Nb de doc trouvés
Carte Id (NdG)	29	83	82
Passeport (NdG)	3	6	6

La figure 17 montre des exemples de détection de cartes d'identité multiples dans une même image.

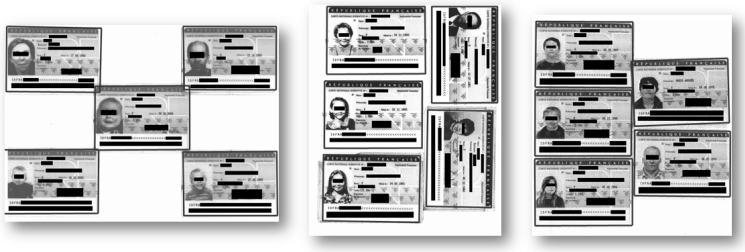


Figure 17. Exemples de multidétection de cartes d'identité. Toutes les cartes sont retrouvées

La figure 18 illustre la qualité de notre proposition en termes de multidétection de modèles de documents différents de celui des cartes d'identité.

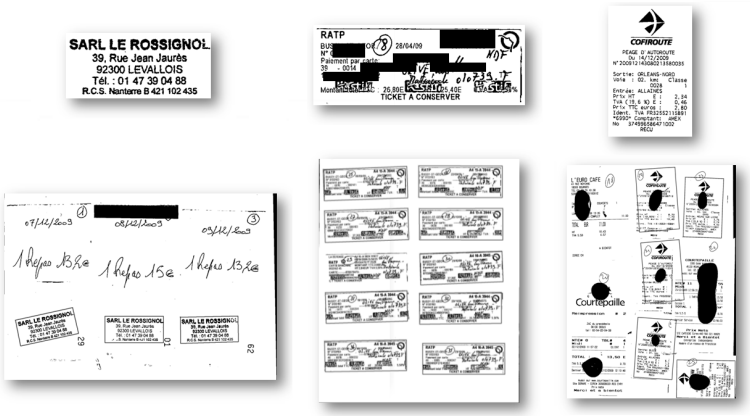


Figure 18. Exemples de multidétection de documents divers. Détection d'un tampon, d'un ticket de métro et d'un ticket d'autoroute. Le ticket de métro non retrouvé n'a pas la même mise en page que les autres tickets

5. Conclusion

Cet article détaille comment il est possible d'adapter au contexte de la comparaison d'images de documents, des techniques utilisées dans le cadre de la recherche par l'exemple d'images naturelles. Les principales adaptations permettent de prendre en compte les fréquentes répétitions de motifs présents dans les images de documents (textes, tableaux...) et les dégradations couramment rencontrées. Les tests mettent en évidence que ces adaptations rendent robuste l'étape de comparaison de points d'intérêt entre images de documents.

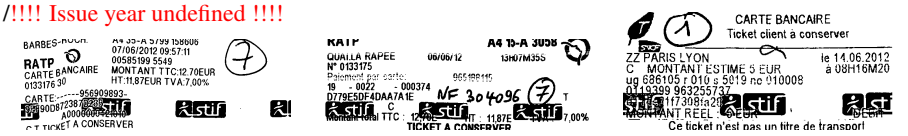
Le second apport de notre proposition est de pouvoir garantir une excellente précision pour l'identification d'images de documents. Des tests effectués en production, à l'aide d'une société de numérisation, nous ont permis d'obtenir un retour sur la qualité de notre méthode dans un cas réel d'utilisation. Ces tests en production ont notamment fait ressortir qu'il était possible, en faisant varier simplement 2 seuils, d'adapter notre proposition à un grand nombre de documents semi-structurés différents.

La principale perspective de ce travail est d'étendre notre proposition à des documents moins structurés que ceux que nous traitons ici. En effet si une classe est composée de documents dont la structure change significativement d'une image à l'autre, notre méthode éprouvera des difficultés à correctement les identifier. Ainsi, pouvoir comparer et identifier des classes de documents telles que celles présentées figure 19, nécessite d'adapter notre méthode.

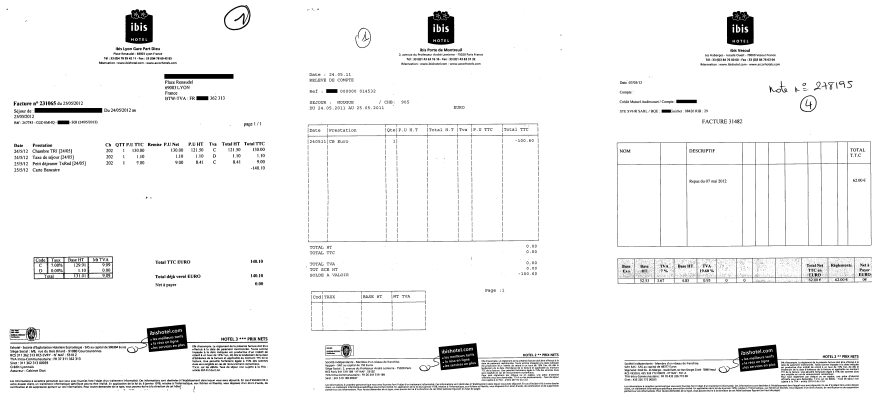
Pour cela, nous travaillons actuellement à l'utilisation des sacs de mots visuels (*bags of visual words*) (J. Yang *et al.*, 2007) afin de permettre la comparaison de documents présentant des motifs communs mais disposés de manière différentes sur les images. En effet, ces outils permettent de ne pas prendre en compte l'information de position des points d'intérêt. Seule la distribution des motifs est prise en compte.

La deuxième perspective de ces travaux est dévaluer l'impact d'un changement des différentes briques composant notre chaine de traitements. Nous pensons par exemple tester les performances d'autres détecteurs et descripteurs de points d'intérêt. Il pourrait être également intéressant de réfléchir à la pertinence de créer des détecteurs ou descripteurs dédiés aux images de documents à l'image de ce qui est fait par les auteurs de (T.-A. Pham *et al.*, 2012). De même, nous envisageons d'apporter les modifications faites à RANSAC à ses variantes LO-RANSAC et SCRAMSAC.

La dernière perspective de ces travaux est liée à l'optimisation des temps de traitement. Si l'extraction des points d'intérêt est rendue plus rapide grâce à l'utilisation de SURF plutôt que SIFT, cette étape peut néanmoins prendre de 3 à 10 secondes par image. Pour cette raison, l'ensemble des points d'intérêt des documents des bases à analyser sont extraits par un traitement hors-ligne. Ceci permet de diminuer (en moyenne) les temps de traitement de 2 secondes par image lors de l'utilisation de notre système en production. Les traitements pourraient également être accélérés en utilisant des descripteurs plus légers comme ceux utilisés dans (Takeda *et al.*, 2011). L'objectif serait d'extraire moins de points, en quantifiant les vecteurs de caractéris-



a) Tickets RATP



b) Factures IBIS

Figure 19. Exemples de deux classes de documents ("RATP" et "IBIS") dont les différents documents présentent de grandes variations visuelles. Ces classes ont une forte variabilité intra-classe. Une portion des documents reste tout de même similaire d'un exemplaire à un autre

tiques ou encore, en développant la sélection de points d'intérêt telle que nous l'avons présentée dans la partie 3.1. Une autre perspective permettant d'optimiser les temps de traitements serait d'utiliser les adaptations de RANSAC citées dans (Raguram *et al.*, 2008).

Remerciements

Les auteurs de l'article remercient l'entreprise de dématérialisation Gestform, pour le temps passé à nous expliquer leur problématique et à nous fournir des images de documents et particulièrement M. Jean-Marc Nahon directeur des études informatiques de l'entreprise. Nous remercions également Philippe Larroye, élève à l'ENSEIRB-MATMECA qui a participé, dans le cadre d'un stage, à la mise au point de l'algorithme.

Bibliographie

Audithan S., Chandrasekaran R. (2009). Document text extraction from document images using haar discrete wavelet transform. *European Journal of Scientific Research*, vol. 36, nº 4, p. 502–512.

Augereau O., Journet N., Domenger J.-P. (2011). Document images indexing with relevance feedback: an application to industrial context. In *Document analysis and recognition (icdar), 2011 international conference on*, p. 1190–1194.

Augereau O., Journet N., Domenger J.-P. (2012). Reconnaissance et extraction de pièces d'identité. In *Actes du douzième colloque international francophone sur l'écrit et le document (cifed)*, p. 179–194.

Bay H., Ess A., Tuytelaars T., Van Gool L. (2008). Speeded-up robust features (surf). *Computer Vision and Image Understanding*, vol. 110, n° 3, p. 346–359.

Bhatti N., Hanbury A. (2011). Local primitive histograms for patent binary image retrieval. In *9th iapr international workshop on graphics recognition (grec)*.

Brown M., Lowe D. (2007). Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, vol. 74, n° 1, p. 59–73.

Cao F. (2008). *A theory of shape identification* (vol. 1948). Springer Verlag.

Chen N., Blostein D. (2007). A survey of document image classification: problem statement, classifier architecture and performance evaluation. *International Journal on Document Analysis and Recognition*, vol. 10, n° 1, p. 1–16.

Chum O., Matas J., Kittler J. (2003). Locally optimized ransac. *Pattern Recognition*, p. 236–243.

Fischler M., Bolles R. (1981). 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.

Harris C., Stephens M. (1988). A combined corner and edge detector. In *Alvey vision conference*, vol. 15, p. 50.

Jain R., Doermann D. (2012). Logo retrieval in document images. In *Document analysis systems (das), 2012 10th iapr international workshop on*, p. 135–139.

Juan L., Gwun O. (2009). A comparison of sift, pca-sift and surf. *International Journal of Image Processing (IJIP)*, vol. 3, n° 4, p. 143–152.

Kalia R., Lee K.-D., Samir B., Je S.-K., Oh W.-G. (2011). An analysis of the effect of different image preprocessing techniques on the performance of surf: Speeded up robust features. In *Frontiers of computer vision (fcv), 2011 17th korea-japan joint workshop on*, p. 1–6.

Ke Y., Sukthankar R. (2004). Pca-sift: A more distinctive representation for local image descriptors.

Lowe D. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, vol. 60, n° 2, p. 91–110.

Matas J., Chum O., Urban M., Pajdla T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, vol. 22, n° 10, p. 761–767.

Mikolajczyk K., Schmid C. (2005). A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, n° 10, p. 1615–1630.

Morel J., Yu G. (2009). Asift: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, vol. 2, n° 2, p. 438–469.

Muja M., Lowe D. (2009). Fast approximate nearest neighbors with automatic algorithm configuration. In *International conference on computer vision theory and applications (vissapp 09)*, vol. 340, p. 331–340.

Nayef N., Breuel T. M. (2013). On the use of geometric matching for both: Isolated symbol recognition and symbol spotting. In *Graphics recognition. new trends and challenges*, p. 36–48. Springer.

Pham T.-A., Delalandre M., Barrat S., Ramel J.-Y. (2012). A robust approach for local interest point detection in line-drawing images. In *Document analysis systems (das), 2012 10th iapr international workshop on*, p. 79–84.

Pham T. D. (2003). Unconstrained logo detection in document images. *Pattern recognition*, vol. 36, n° 12, p. 3023–3025.

Psyllos A., Anagnostopoulos C., Kayafas E. (2010). Vehicle logo recognition using a sift-based enhanced matching scheme. *Intelligent Transportation Systems, IEEE Transactions on*, vol. 11, n° 2, p. 322–328.

Raguram R., Frahm J., Pollefeys M. (2008). A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus. *Computer Vision–ECCV 2008*, p. 500–513.

Rangoni Y., Belaïd A. (2012). Segmentation de documents composites par une technique de recouvrement des espaces blancs. In *Cifed-coria.*

Rusinol M., Lladós J. (2009). Logo spotting by a bag-of-words approach for document categorization. In *2009 10th international conference on document analysis and recognition*, p. 111–115.

Sattler T., Leibe B., Kobbelt L. (2009). Scramsac: Improving ransac's efficiency with a spatial consistency filter. In *Computer vision, 2009 ieee 12th international conference on*, p. 2090–2097.

Silpa-Anan C., Hartley R. (2008). Optimised kd-trees for fast image descriptor matching. *Computer Vision and Pattern Recognition*, p. 1–8.

Sur F., Noury N., Berger M.-O. (2011, July). *Image point correspondences and repeated patterns*. Research Report n° 7693. INRIA.

Takeda K., Kise K., Iwamura M. (2011). Real-time document image retrieval for a 10 million pages database with a memory efficient and stability improved llah. , p. 1054–1058.

Tuytelaars T., Mikolajczyk K. (2008). Local invariant feature detectors: a survey. *Foundations and Trends in Computer Graphics and Vision*, vol. 3, n° 3, p. 177–280.

Uchiyama H., Saito H. (2009). Augmenting text document by on-line learning of local arrangement of keypoints. In *Mixed and augmented reality, 2009. ismar 2009. 8th ieee international symposium on*, p. 95–98.

Viola P., Jones M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer vision and pattern recognition, 2001. cvpr 2001. proceedings of the 2001 ieee computer society conference on*, vol. 1, p. I–511.

Yang G., Stewart C., Sofka M., Tsai C. (2007). Alignment of challenging image pairs: Refinement and region growing starting from a single keypoint correspondence. *IEEE Trans. Pattern Anal. Machine Intell*, vol. 23, n° 11, p. 1973–1989.

Yang J., Jiang Y., Hauptmann A., Ngo C. (2007). Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on workshop on multimedia information retrieval*, p. 197–206.

Zhu G., Doermann D. (2009). Logo matching for document image retrieval. In *Document analysis and recognition (icdar), 2009 international conference on*, p. 606–610.