Université de Paris

École doctorale Informatique, Télécommunications et Électronique (EDITE ED130)

Laboratoire Informatique Paris Descartes (LIPADE EA2517)

Système Intelligent de Perception (SIP)

Extraction de la mise en page de documents :

Application à la sécurisation des documents hybrides

Par Héloïse ALHERITIERE

Thèse de doctorat

Spécialité : Informatique

Présentée et soutenue publiquement le [Date de soutenance]

Jury :

Directeurs de thèse	M^{me} Nicole VINCENT	Professeur, Université de Paris
	M. Jean-Marc OGIER	Professeur, Université de La Rochelle
Rapporteurs	M. Antoine TABBONE	Professeur, Université de Lorraine
	M^{me} Véronique EGLIN	Professeur, INSA de Lyon
Examinateurs	M^{me} Laurence LIKFORMAN	Maître de conférence (HDR), Télécom Paris Tech
	M. Jean-Yves RAMEL	Professeur, Université de Tours
	M^{me} Florence CLOPPET	Maître de conférence (HDR), Université de Paris
	M. Camille KURTZ	Maître de conférence, Université de Paris
Invitée	M^{me} Petra GOMEZ-KRAMER	Maître de conférence, Université de La Rochelle

Résumé

Les documents numériques sont de plus en plus présents dans notre société. Ce format a de nombreux avantages, que ce soit pour la diffusion ou la sauvegarde de documents. La diffusion permet de transmettre facilement des documents, mais ne permet pas de garantir l'intégrité de ceux-ci, ni pour ceux qui le reçoivent, ni pour ceux qui le diffusent. Durant leur cycle de vie, les documents passent généralement d'un état dématérialisé à un état matérialisé et inversement. Les deux formats possèdent leurs avantages et leurs inconvénients, ce qui justifie qu'un même document puisse se retrouver dans les deux états. Lorsque l'on passe d'un format matérialisé à celui dématérialisé, nous obtenons une image, un ensemble de pixels qu'il faut interpréter. Les différentes instances d'un même document que nous pouvons obtenir en scannant ou en imprimant plusieurs fois celui-ci définissent le « document hybride ».

Un premier niveau de comparaison peut être réalisé en analysant la mise en page du document. Les méthodes d'extraction de la mise en page sont nombreuses et nous les analysons pour mettre en évidence leurs défauts et leur adéquation à des catégories bien particulières de document. Aussi nous avons développé une méthodologie qui s'appuie sur de nouvelles transformées permettant d'innover dans le mode de représentation d'une image de document. Les segments de droites sont au centre de notre travail. Nous pouvons traiter des documents divers sans avoir recours à un apprentissage supervisé. Nous innovons aussi au niveau de l'évaluation de notre proposition. En effet, dans la perspective de la sécurisation d'un document hybride, à la précision d'une décomposition de la page, nous adjoignons la nécessité de résultats stables pour toutes les instances d'un document.

Abstract

Digital documents are more and more present in our society. This format has many advantages, whether for distribution or document backup. Distribution allows for an easy transmission of documents but do not guarantee their integrity neither for the receiver nor for the sender. Throughout their life cycle, documents go from a dematerialized state to a materialized state and *vice versa*. The two formats have their own advantages and disadvantages, justifying the fact that a document can be found in the two formats. When we go from a materialized format to a dematerialized one we get an image, a set of pixels that need to be interpreted. The different instances of a same document obtained by scanning or printing it many times define the "hybrid document".

A first level of comparison can be realized by analyzing the document layout. Many layout extraction methods exist. We analyze them to highlight their default and their adequacy to particular category of documents. We have also developed a methodology based on new transforms thus innovating in the representation of a document image. We can process various documents without needing supervised learning. We also adopt a more innovative approach in our evaluation method. Thus, for the purpose of securing hybrid document, we associate to the accuracy of a page decomposition the necessity of stable results for every instance of a document.

Table des matières

Table des figures

Liste des tableaux

Introduction

De plus en plus fréquemment nous sommes amenés à devoir transmettre des documents. Ces échanges sont utilisés pour partager des informations, pour fournir des renseignements de nature administrative ou des justificatifs de toutes sortes. Ces documents échangés peuvent prendre différentes formes en fonction des différents supports sur lesquels ils se présentent : papier lorsqu'ils apparaissent sous forme physique ou fichiers numériques lorsqu'ils sont stockés par exemple dans nos ordinateurs (ou tablettes, smartphones, cartes, badges, *etc.*). La dématérialisation des documents « papier » est omniprésente dans notre société. Cette étape est de plus en plus fréquente dans la vie du document. Dans notre monde hyper connecté, pratiquement toutes les informations circulent par voie électronique. En effet, un document numérique possède de nombreux avantages. Il peut être envoyé à l'autre bout de la planète quasi-instantanément, facilitant largement sa diffusion. Par ailleurs, au moment de la création du document, la constitution et la mise en page sont extrêmement facilitées. Une erreur de typographie est très facile à corriger. C'est là un avantage certain par rapport aux anciens modes de production comme la rédaction manuscrite ou l'usage de la machine à écrire.

Mais cet avantage devient un problème lors de l'échange ou de l'utilisation d'un document dont nous ne sommes pas l'auteur. Savoir d'où il vient, qui l'a produit, renforce sa crédibilité. La notion de confiance intervient alors et impose parfois un protocole adapté. La sécurisation des documents est une problématique aussi vieille que l'invention de l'écriture. Chaque époque a innové pour sécuriser les documents avec les moyens dont elle disposait. Aujourd'hui, la sécurisation des documents est une préoccupation croissante. En effet, les moyens disponibles pour falsifier les documents sont de plus en plus performants, ce qui implique que les moyens de sécurisation doivent être de plus en plus efficaces. Cette sécurisation

répond à deux nécessités : garantir que le document n'a pas été altéré par rapport au document original et/ou protéger le document pour que seules des personnes habilitées puissent le lire.

Assurer la fidélité d'un contenu permet de lutter contre la fraude en permettant de rejeter un document non fidèle. Comme nous le rappelions plus haut, cette irruption des documents numériques envahit notre quotidien. De nombreux exemples peuvent être relatés. Par volonté d'économie, les banques n'envoient presque plus de relevés de compte par courrier postal. De manière générale, de plus en plus de sociétés ont recours à l'envoi dématérialisé de leurs factures. Ainsi, après la FSE (feuille de soins électronique qui représente 90% du traitement des remboursements pour les complémentaires), la Sécurité Sociale lance un grand projet de dématérialisation. Et même si l'administration française demande encore la version papier des documents, il s'agit de plus en plus de documents que les usagers ont dû scanner et imprimer. Ce processus rend caducs certains moyens existants pour sécuriser les documents « papier » comme le filigrane. Dans le même temps, le gouvernement rappelle que fabriquer et / ou utiliser un faux document sont des délits punissables de 3 ans d'emprisonnement et de 45 000 € d'amende. Ces peines sont aggravées si le faux document est un document délivré habituellement par une administration (faux papiers d'identité, fausse carte Vitale...) passant la peine à 5 ans de prison et à 75 000 € d'amende. Le simple fait de détenir des faux documents est puni par 2 ans de prison et 30 000 € d'amende.

Un document peut donc avoir plusieurs états : papier ou numérique. Cette propriété peut être native ou obtenue par différents processus. Le passage de la version papier à la version numérique se fait par acquisition d'une image composée d'un ensemble de pixels. Cette acquisition peut être réalisée par un scanner, mais pas uniquement. La démocratisation et l'amélioration des appareils photos, notamment ceux de nos téléphones portables, permettent désormais la numérisation de documents par ceux-ci. Grâce à ces nouveaux moyens, l'utilisateur gagne en temps et en simplicité pour numériser un contenu bien que cela entraîne une complexité supplémentaire pour les traitements automatiques qui seront ultérieurement appliqués à l'image du document. Dans l'état actuel de la technologie, le passage de la version numérique à la version papier est réalisé grâce à une imprimante. Celles-ci ne restituent pas toujours le document de la même façon. La différence est très marquée entre les imprimantes à jet d'encre et les imprimantes laser. Mais elles ne sont pas seules en cause. En effet, chaque cartouche d'encre possède sa propre teinte et les résultats peuvent donc en dépendre. Les moyens de sécurisation ne sont donc pas les mêmes, ils varient en fonction du support du document.

Une façon d'interpréter un document consiste à en extraire la mise en page. On entend par là l'ordonnancement et l'organisation spatiale des différents éléments constituant un document. Il s'agit à la fois d'identifier les éléments qui le composent mais également de structurer celui-ci. En effet, le passage à une image, ensemble de pixels, a conduit à la perte des structurations du document construit par son auteur. La mise en page aide ainsi à la compréhension et au développement des idées. Si un document possède une mauvaise mise en page, cela rend généralement le document incompréhensible ou difficilement compréhensible. Une phase de réflexion est alors nécessaire au lecteur.

L'extraction de la mise en page revient donc à identifier les éléments qui composent un document et leur localisation. Les éléments sont généralement du texte, des tableaux, des images, des logos, *etc.* À cela, on peut ajouter le complémentaire de ces éléments : le fond. Ce sont donc des éléments assez différents les uns des autres que l'on distingue sans ambiguïté. Néanmoins, nous pouvons remarquer qu'un tableau et du texte ne sont pas exhaustifs l'un de l'autre ; en effet un tableau est généralement composé de texte. Pour autant, un tableau possède des caractéristiques qui lui sont propres tels les alignements ou les séparateurs qui rendent plus aisée la lecture.

Ainsi, nous cherchons à extraire et à identifier les éléments qui composent un document. L'identification du type des éléments comme texte, image, logo, *etc.* permet de pouvoir comparer deux éléments entre eux, en ne prenant en compte que des caractéristiques invariantes à une dégradation naturelle. C'est cette comparaison qui contribue à sécuriser les documents. En effet, comparer deux éléments qui sont localisés au même emplacement et qui sont issus d'instances d'un même document hybride permet de savoir s'il s'agit vraiment du même document. Nous considérons dans le cadre de cette thèse le terme de « stabilité » pour nommer ce qui permet de mesurer la répétition des résultats par rapport à des instances différentes. Ainsi, si nous considérons un simple document composé par exemple de deux paragraphes, un résultat stable sur plusieurs instances de ce document devra donner toujours deux paragraphes comportant les mêmes mots, peu importe si les images sont différentes en termes de pixels (qu'elles aient subit une translation ou une rotation par exemple).

Contexte

Cette thèse a été financée dans le cadre du projet pluridisciplinaire (droit et informatique) labélisé par l'Agence Nationale de la Recherche (ANR) intitulé « Hachage sémantique pour la signature électronique avancée de document *(Semantic Hash for Advanced*

Document Electronic Signature (SHADES)) »[1]. Ce projet associe plusieurs laboratoires et un partenaire industriel. Plus précisément deux laboratoires d'informatique : le Laboratoire d'Informatique PAris DEscartes (LIPADE) et le Laboratoire Informatique, Image et Interaction (L3i), ainsi que le Centre d'Études Juridiques et Politiques (CEJEP), la fédération des Tiers de confiance du Numérique (FNTC) et l'entreprise ITESOFT. Pour sa part, le CEJEP est chargé des questions juridiques très présentes dans le cadre du projet, l'entreprise ITESOFT, qui a participé à la phase de recherche, a la responsabilité de coordonner les différentes parties et de créer un prototype, les laboratoires d'informatique celle de modéliser une solution fondée sur le traitement d'images. La FNTC joue le rôle du client. La sécurisation de documents pouvant passer de l'état dématérialisé à l'état matériel (et inversement) est l'objet du projet SHADES. Pour ce faire, le document va être analysé selon son contenu et une « signature » va lui être assignée. Cette signature étant fondée sur le contenu, elle doit être identique quelle que soit l'instance du document. L'une des visées initiales du projet était de sauvegarder les « signatures » dans une base de données et de vérifier si le document soumis est le même que l'un de ceux déjà présents dans la base. La réflexion a été poussée jusqu'à savoir si le résultat du projet pouvait être employé devant un jury et affirmer que ce document est le même que celui qui a été déposé dans la base de données à telle date.

Un document évolue lors de son cycle de vie et c'est bien souvent la concordance entre le document initial et son image qui doit donc être confortée. L'image envoyée sera fréquemment altérée. Ainsi, elle peut varier en fonction de sa résolution, présenter une couleur différente du document original en fonction du réglage de l'imprimante utilisée. Elle peut également présenter une bordure supplémentaire ou être parsemée de taches provenant de la présence de poussières sur la vitre du scanner. Tout cela ne modifie pourtant pas, à nos yeux, le document. Dans ce contexte, il est difficile de dire quel est le « véritable document » qui fait foi. Ainsi, nous n'échangeons souvent qu'une copie du document initial. Si on scanne à nouveau l'image imprimée, alors l'image obtenue sera probablement différente de l'image du document précédemment imprimé. Une comparaison pixel à pixel conduirait, dans tous les cas, à des différences qui pourraient laisser penser que le document a été modifié alors que ce n'est pas le cas. La sécurisation de l'image du document impose donc de mieux cerner le contenu du document et de limiter les comparaisons à ce contenu en sachant différencier les déformations naturelles des déformations malveillantes. Pour traiter ce problème, dans le projet SHADES, nous avons introduit la notion de document hybride (cf. Figure 1) : toutes les copies du document ainsi que le document original sont considérés comme constituant un unique

[1] ANR-14-CE28-0022.

document hybride. Et cela, évidemment, à condition que les copies puissent encore être considérées légalement comme le même document.

Il est donc, dans un premier temps, important de préciser ce qu'est un document. Traditionnellement, il s'agit d'une pièce écrite donnant des renseignements divers ou servant de preuve, de témoignage. Dans ce cadre, le document papier a une existence légale. Pour ce qui est du document numérique, apparu plus récemment, seule la signature, si elle répond à un protocole sécurisé, a une valeur légale.

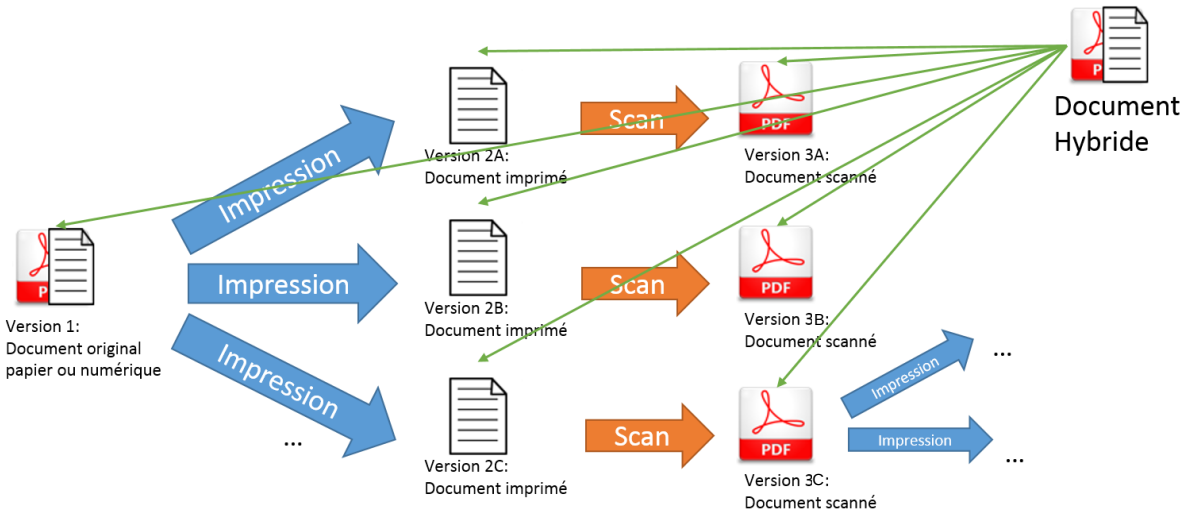


Figure 1 - Illustration d'une portion d'un cycle de vie d'un document hybride.

La problématique de la sécurisation est ainsi partiellement liée à la comparaison de deux documents. Il est alors nécessaire de comprendre comment un observateur perçoit et peut décrire un document. De manière concrète, le premier niveau d'observation d'un document consiste à en extraire les grandes parties ainsi que l'organisation de ces parties. Autrement dit, la mise en page d'un document permet d'en identifier la nature et la localisation du contenu. En second niveau, on peut analyser la suite des caractères dans une zone de texte ou la répartition des couleurs dans une image. Nous allons centrer notre propos sur l'identification de la structure du document qui est la première étape dans un processus de sécurisation du document.

Problématiques

Dans ce contexte, l'objectif de cette thèse est d'extraire la mise en page dans des images de documents contemporains afin de les sécuriser. Néanmoins, notre approche doit rester

cohérente avec tous types de documents, ce qui constitue déjà un premier verrou méthodologique.

Par ailleurs, nous nous efforcerons d'extraire la mise en page de documents hétéroclites de façon stable, ce qui est le fil conducteur de ces travaux. Comme nous le verrons, il existe plusieurs façons d'extraire la mise en page, posant différents problèmes. La segmentation, qui permet de diviser le document en un certain nombre de régions, est une des étapes possibles. La segmentation d'images est un problème complexe qui est mal posé car il existe une infinité de résultats / partitions possibles. Un « bon » résultat est un résultat qui répond au problème donné dans l'application. Généralement lorsque l'on traite de segmentation, l'axe de recherche est celui de l'amélioration des résultats par la qualité. La question de la stabilité est une perspective nouvelle qui soulève de nombreuses questions.

Le problème principal dans cette démarche et dans le cadre du projet SHADES est l'obtention d'un résultat de segmentation stable. Cette stabilité pose notamment la question de l'incertitude que l'on est prêt à prendre en compte. En effet, une stricte stabilité supposerait d'avoir strictement le même nombre de régions et les mêmes positions.

La segmentation suppose bien souvent une post-étape de labélisation (étiquetage) des régions permettant de reconnaître le type de celles-ci (par exemple en texte, image, *etc.*). Comme pour la segmentation, l'étape de labélisation doit être regardée sous le prisme de la stabilité, ce qui est également un problème nouveau à considérer.

Dans le cadre du projet et pour traiter un maximum de documents avec des mises en page très différentes, nous avons choisi de ne pas utiliser d'apprentissage supervisé qui s'appuierait sur les caractéristiques des documents et donc aurait une capacité de généralisation limitée. Concevoir un tel système automatique, sans apprentissage, constitue également une problématique étudiée dans ces travaux de thèse.

Organisation du mémoire

Ce manuscrit de thèse est structuré en quatre chapitres. Le premier chapitre présente un état de l'art sur les méthodes d'extraction de la mise en page en analysant les images de documents. Le second chapitre présente les nouvelles transformées que nous avons définies pour construire une méthodologie d'extraction de la mise en page. Cette méthodologie sera détaillée dans le chapitre 3. Nous présenterons dans le chapitre suivant le moyen de sécuriser les documents hybrides pour finir par une évaluation de notre méthode d'extraction de la mise en

page selon le critère permettant de garantir la stabilité. Le manuscrit se termine par une conclusion dans laquelle nous aborderons des perspectives possibles à notre travail.

Chapitre 1 : Analyse de la mise en page

L'analyse de la mise en page est un domaine très étudié dans la problématique du traitement des documents. En effet, elle regroupe la détection, la localisation et l'identification des différents éléments composant le document. De manière à mieux positionner notre proposition, nous dressons dans ce chapitre un état de l'art sur les différentes méthodes d'extraction de la mise en page. Nous commencerons par les méthodes de segmentation qui permettent de diviser les documents en régions homogènes, puis celles utilisant une classification des pixels pour extraire la mise en page. Enfin, nous présenterons les méthodes extrayant successivement des couches spécifiques. Pour chaque méthode, une présentation de celle-ci avec les avantages et les inconvénients sera donnée.

Chapitre 2 : L'approche par les lignes : de nouvelles transformées

Dans ce chapitre, nous proposons de nouvelles transformées fondées sur les lignes et permettant de décrire une image en fonction d'informations de longueur ou d'orientation. Ces transformées peuvent être définies globalement ou localement en fonction de l'utilisation qui nous intéresse. Nous présenterons également les propriétés que possèdent ces transformées qui ont été définies dans le domaine du continu. De plus, nous présenterons comment ces transformées peuvent être efficaces dans le domaine du discret, ce domaine correspondant au domaine de définition des images.

Chapitre 3 : Extraction de la mise en page dans les documents

Dans ce chapitre, nous présenterons notre méthode pour extraire la mise en page fondée sur la dualité entre le fond et la forme que nous décrivons en utilisant les transformées que nous avons définies dans le chapitre précédent. Dans une première étape, l'extraction des séparateurs matérialisés conduit à un premier découpage de parties telles les tableaux et permet alors une simplification de la tâche. La deuxième étape utilise les séparateurs implicites pour segmenter le document en régions. Enfin, la dernière étape labélise ces régions. Nous évaluerons les différentes étapes selon le critère de qualité des résultats.

Chapitre 4 : Sécurisation des documents hybrides

Dans ce chapitre, nous nous intéressons à la sécurisation des documents hybrides. Après un rapide aperçu des méthodes portant sur ce thème, nous présenterons le système de

sécurisation des documents du projet SHADES en le positionnant par rapport aux méthodes de sécurisation existantes. Dans ce chapitre, nous aborderons et préciserons des notions en lien avec la sécurisation, telles que l'égalité, la robustesse et la stabilité d'un algorithme. Cela nous conduira à introduire un nouveau point de vue pour l'évaluation des résultats. Ainsi la méthode exposée au chapitre 3 sera estimée en fonction de ces derniers critères : la stabilité et la robustesse.

Chapitre 1
Analyse de la mise en page

Sommaire

Résumé

L'analyse de la mise en page est un domaine très étudié dans la problématique du traitement des documents. En effet, elle regroupe la détection, la localisation et l'identification des différents éléments composant le document. De manière à mieux positionner notre proposition, nous dressons dans ce chapitre un état de l'art sur les différentes méthodes d'extraction de la mise en page. Nous commencerons par les méthodes de segmentation qui permettent de diviser en régions homogènes les documents, puis celles utilisant une classification des pixels pour extraire la mise en page. Enfin, nous présenterons les méthodes extrayant successivement des couches spécifiques. Pour chaque méthode, une présentation de celle-ci avec les avantages et les inconvénients sera donnée. Nous proposerons plusieurs tableaux récapitulatifs permettant une comparaison des méthodes.

1.1 Introduction

L'analyse de la mise en page (dit *layout* en anglais) d'un document permet d'identifier les différentes parties composant celui-ci et leurs agencements. Cette analyse est importante pour la dématérialisation et la lecture de ce dernier sur de nouveaux supports comme un téléphone mobile ou une tablette. Ces nouveaux supports ont la particularité d'avoir de petits écrans. La structure d'un document n'est pas conservée dans la structure de l'image numérisée qui est limitée à une matrice de pixels. L'objectif est alors de la reconstituer au moins partiellement. L'analyse de la mise en page peut également servir à la classification d'images de documents. Elle permet de classifier les types de documents selon différentes catégories (comme lettres, factures, articles, *etc.*) sans avoir d'*a priori* sur le contenu de la base d'images à traiter [Auge13]. La mise en page, pour structurer le contenu d'un document, peut être considérée à différents niveaux (cf. Figure 2) :

- La mise en page **physique** correspond à ce qui est identifiable sans connaître la langue du contenu textuel ou la culture des rédacteurs qui influence bien sûr la création du document. Elle comprend, en elle-même, différents aspects. L'aspect géométrique correspond à la perception de larges zones homogènes comme les textes, les images ou les graphiques. Mais également, au sein des textes, on perçoit les colonnes distinctes ou les paragraphes. Une vision plus rapprochée permet de distinguer des éléments de typographie que l'émetteur du document a introduit pour mettre en valeur certains points de son message (italique, gras, souligné, *etc.*). La taille des caractères a aussi son importance dans la compréhension du document ;

- La mise en page **logique** s'enrichit des informations à caractère sémantique de titre, résumé, pied de page, section, *etc.* qui organisent, structurent et hiérarchisent le contenu d'un document. Cet aspect de la mise en page constitue l'interprétation que les auteurs ont voulu donner à une partie de ce document. L'analyse de la mise en page logique permet également de reconstituer l'ordre de lecture d'un document, c'est-à-dire l'ordre naturel dans lequel nous devrions lire les différents éléments du document afin d'en optimiser la compréhension.

Dans le cadre du projet SHADES, une analyse logique du document n'apportera pas d'information supplémentaire pertinente pour créer une signature du document. En effet, même si le texte est un pied de page, il peut contenir une information importante et doit être traité avec la même importance qu'un titre ou une unité de lecture. Par conséquent, nous n'aborderons pas cet aspect dans la suite de ce document. Nous ne nous intéresserons, dans ce manuscrit, qu'à l'aspect physique de la mise en page.

(a) (b)

Figure 2 - Illustration des différents niveaux de mise en page d'un document. (a) Document initial. (b) Représentation des différents labels (en rouge, la mise en page logique et en noir, le mise en page physique).

Parmi les approches possibles pour extraire la mise en page physique d'une page, certaines requièrent préalablement de segmenter l'image avant de labéliser les zones. Une méthode de segmentation d'images est une méthode qui partitionne l'image en un ensemble de composantes connexes homogènes. On pourra distinguer d'une part les méthodes traditionnelles de traitement d'images qui en fonction d'un critère d'homogénéité rassemblent les pixels en zones, et d'autre part les méthodes qui classifient les pixels en fonction de caractéristiques.

La multitude des documents possédant des contenus très hétéroclites et l'hétérogénéité entre ces différents éléments sont les principales difficultés de l'analyse de la mise en page. Si l'on compare les documents anciens aux documents récents, ils n'ont généralement pas les mêmes caractéristiques. Les documents les plus anciens traités ont été écrits avant la naissance de l'imprimerie. Ils sont donc constitués de textes manuscrits, de lettrines et d'images peintes à la main. Ils souffrent également de problèmes de conservation. Ainsi, l'écriture a pu être effacée, le papier changer de couleur en vieillissant. Les documents contemporains, eux, peuvent avoir des mises en page plus complexes dues à l'amélioration de nos outils de production rendant ainsi les documents plus esthétiques mais plus difficiles à analyser. La Figure 3 présente une variété de documents récents où, en rouge, sont encadrés les différents éléments de la mise en page. Les documents ont été choisis dans la base *PRImA Layout Analysis Dataset* [2] qui fait depuis 2003 régulièrement l'objet de compétitions pour l'extraction de la mise en page, organisées à l'occasion de la conférence internationale ICDAR *(International Conference on Document Analysis and Recognition)* [ClAP17, ACPP15, ACPP13, ACPP11, AnGB07, AnGB05, AnGK03].

[2] https://www.primaresearch.org/dataset/.

Figure 3 - Exemples de mise en page de documents extraits de la base PRImA Layout Analysis Dataset. Cette image est extraite du site de la base de données contenant une vérité terrain (physique et logique).

Le traitement de cette typologie hétéroclite des documents a fait l'objet d'un état de l'art exhaustif [Trup05] abordant différentes méthodes d'analyse d'images de documents allant des prétraitements à utiliser à la reconnaissance des blocs, en passant par l'analyse de l'ordre de lecture. Cet état de l'art est l'un des seuls qui aborde également les différents types de données. Les différentes méthodes sont brièvement abordées. C'est pourquoi nous détaillerons davantage, dans la suite de ce chapitre, certaines de ces méthodes et les compléterons par les études qui ont été réalisées depuis cette date.

L'étude de Journet *et al.* [JMER07] présente un état de l'art sur l'analyse des documents anciens. Ces documents ont des problématiques différentes de ceux que nous voulons traiter. Néanmoins, nous pouvons parfois reprendre les mêmes approches, dans la mesure où l'hétérogénéité des pages composant le corpus d'images de documents anciens impose une remise en cause des outils traditionnels d'analyse.

Certaines méthodes dont nous parlerons dans ce chapitre ont comme inconvénients de ne pas fournir de résultats pertinents sur des images de documents inclinés ou d'être trop sensibles au bruit. Ce sont des défauts qu'un prétraitement pourrait corriger (même si cela peut entraîner une perte de qualité dans le cas du redressement de l'image). Il faut toutefois faire attention car un prétraitement peut également éliminer quelques éléments considérés à tort comme du bruit par la méthode choisie. Ce risque est un écueil majeur étant donné que l'un des buts du projet SHADES est de fournir une preuve devant un jury.

Nous aborderons en premier lieu, dans ce chapitre, les méthodes qui aident à extraire la mise en page par une segmentation des documents (cf. Section 1.2). Puis nous évoquerons les méthodes de classification (cf. Section 1.3) et ensuite les méthodes ne recherchant dans les images qu'un seul type de media (cf. Section 1.4). Nous finirons par une synthèse et une discussion (cf. Section 1.5).

1.2 État de l'art sur les méthodes de segmentation des images de documents

Une méthode de segmentation d'images est une méthode qui partitionne l'image en un ensemble de composantes connexes homogènes. Cela peut se faire par un regroupement de pixels constituant une zone répondant aux mêmes propriétés ou par la recherche de frontière qui délimiterait les différents éléments se trouvant dans l'image que l'on veut partitionner. Les partitions trouvées peuvent être différentes en fonction de la définition de la notion d'homogénéité considérée. Par exemple, l'objectif peut être de segmenter l'image de document en paragraphes, en lignes ou en caractères. Après l'étape de segmentation, une caractérisation des régions ainsi trouvées pourra être effectuée pour identifier le label de celles-ci, c'est ce que nous aborderons dans la Section 1.3.4.

On classe généralement les différentes méthodes de segmentation de document en trois catégories : les méthodes descendantes, les méthodes ascendantes et les méthodes hybrides. Les méthodes descendantes, dites « *top-down* » en anglais, partent de l'image globale du document (une partition où tout le support de l'image forme une unique région) qu'elles décomposent récursivement pour, au final, définir des blocs homogènes. Les méthodes ascendantes, dites « *bottom-up* » en anglais, procèdent au contraire par regroupement, en partant de l'analyse des composantes de bas niveau, par exemple les pixels, pour essayer de les regrouper successivement afin de former itérativement des régions de tailles de plus en plus importantes. Les méthodes hybrides combinent ces deux approches. Nous analysons les différents types de méthodes dans les sections suivantes.

1.2.1 Méthodes descendantes

De nombreuses méthodes considèrent le document dans son ensemble pour, récursivement, définir des zones de plus de plus en plus petites. Nous commencerons par présenter une méthode représentative nommée *X-Y cut*.

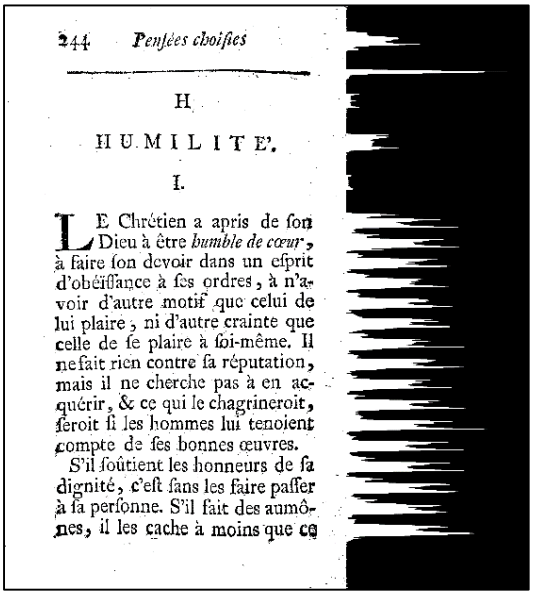
- **Approche X-Y cut**

L'approche « *X-Y cut* » a été proposée dans un premier temps par Nagy et Seth en 1984 [NaSe84]. C'est sans doute la plus ancienne des méthodes consacrées à l'extraction de la mise en page des documents. L'approche repose sur l'hypothèse qu'un texte est constitué de lignes de texte horizontales séparées par un interligne blanc.

Cette méthode s'applique aussi bien sur une image binaire que sur une image en niveaux de gris. Dans le cas d'une image binaire, l'histogramme de projection horizontale indique le nombre de pixels noirs ou blancs (en fonction du fond ou de la forme que l'on cherche à

considérer) par lignes de pixels de l'image (cf. Figure 4) et il résume en une dimension le contenu de l'image. L'histogramme de projection peut également être calculé sur les colonnes permettant d'avoir un histogramme de projection verticale. Dans le cas d'une image en niveaux de gris, la représentation de l'image par l'histogramme de projection contiendra la somme des valeurs des pixels sur chaque ligne. Dans les deux cas, l'histogramme horizontal en une dimension des pixels blancs contient des *extrema* au niveau de chaque interligne. La détection des *extrema* significatifs permet de localiser les interlignes et donc d'en déduire la position potentielle des lignes. La hauteur de ces interlignes est un indicateur de la limite des paragraphes de texte.

Pour segmenter un document, une découpe horizontale ou une découpe verticale dans le cas d'un document multi-colonnes est faite dans un premier temps en considérant les *extrema*. Puis les sous-images sont découpées verticalement de la même façon. Le processus est répété jusqu'à ce qu'il n'y ait plus de coupes possibles. Cette méthode est également appelée *« Recursive X-Y cut »*. Elle est aussi utilisée pour construire un arbre dont les nœuds et les feuilles sont les différents éléments progressivement segmentés. L'arbre X-Y peut également être utilisé pour trouver l'ordre de lecture [Meun05], [Suth10].


244 *Penſées choiſies*

H

HUMILITE'.

I.

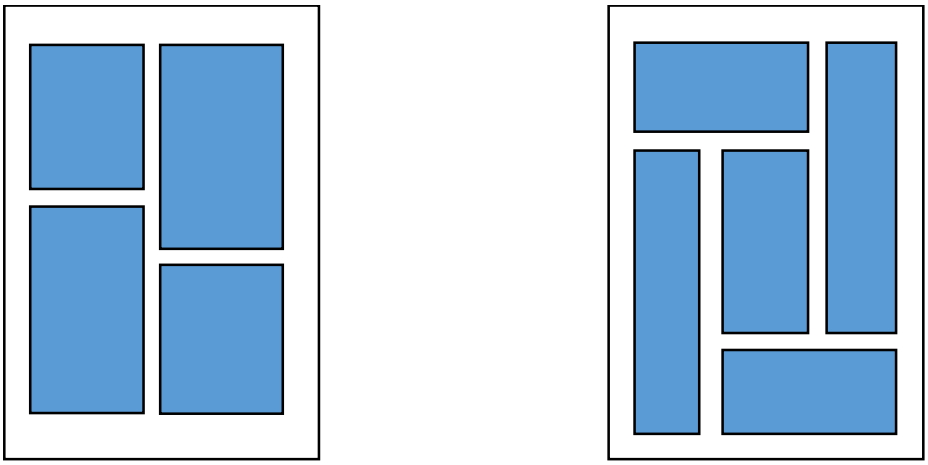
LE Chrétien a apris de ſon Dieu à être *humble de cœur*, à faire ſon devoir dans un eſprit d'obéiſſance à ſes ordres, à n'avoir d'autre motif que celui de lui plaire, ni d'autre crainte que celle de ſe plaire à ſoi-même. Il ne fait rien contre ſa réputation, mais il ne cherche pas à en acquérir, & ce qui le chagrineroit, ſeroit ſi les hommes lui tenoient compte de ſes bonnes œuvres.

S'il ſoûtient les honneurs de ſa dignité, c'eſt ſans les faire paſſer à ſa perſonne. S'il fait des aumônes, il les cache à moins que ce

Figure 4 - Exemple d'histogramme de projection horizontale des pixels noirs (à droite) d'une image binaire de document ancien imprimé (à gauche).

L'un des inconvénients de la méthode *« X-Y cut »* est qu'elle ne détecte que des blocs rectangulaires séparés sur la largeur ou la hauteur de la portion considérée dans le document, ce qui convient aux documents ayant une mise en page « Manhattan », c'est-à-dire rectangulaire comme dans le cas de la Figure 5 (a) mais cette approche ne convient pas pour

des documents ayant des mises en page plus complexes comme illustré sur la Figure 5 (b). Elle n'est pas adaptée lorsque le contenu du document est mal positionné dans l'image, c'est-à-dire lorsque le contenu est incliné.



(a) (b)

Figure 5 - Exemples de types de mises en page. (a) Exemple de mise en page « Manhattan ». (b) Exemple de mise en page non-Manhattan.

Les méthodes descendantes autres que *XY-Cut* ne sont plus utilisées pour l'analyse de la mise en page de documents. Cela est notamment attesté par le fait que les articles récents qui présentent des états de l'art sur les méthodes de segmentation ne présentent plus que la méthode *XY-Cut* dans les méthodes descendantes.

1.2.2 Méthodes ascendantes

Nous présentons maintenant les méthodes dites ascendantes, qui partent de l'ensemble des pixels pour obtenir une segmentation *via* des agrégations successives. Elles reposent généralement sur les lois de la théorie de la Gestalt, auxquelles nous allons consacrer une partie de cette section. Puis nous présenterons les méthodes les plus représentatives : le *Run Length Smoothing Algorithm (RLSA)*, les approches multi-résolutions, la détection des espaces blancs et finalement l'analyse des composantes connexes.

- **La théorie de la Gestalt : psychologie de la forme**

« Die Gestalt » est le nom allemand de la forme. La théorie de la Gestalt prend ses origines dans la psychologie, la philosophie et la biologie [Koff35] permettant de comprendre comment notre cerveau perçoit, analyse et interprète les formes qu'il observe. Elle se divise en six lois formulées par Wertheimer :

La loi de bonne forme : notre cerveau cherche à reconnaître des formes simples et stables qui lui sont familières. Notre perception des éléments se fera de manière globale, en cherchant à regrouper des éléments qui vont ensemble. C'est ainsi que sur le vase de Rubin on peut voir soit les deux visages, soit le vase mais pas les deux en même temps. Cela est illustré sur la Figure 6 (a).

La loi de continuité : des points rapprochés tendent à représenter des formes lorsqu'ils sont perçus. Notre cerveau perçoit d'abord dans une continuité, comme des prolongements les uns par rapport aux autres (cf. Figure 6 (b)).

La loi de la proximité : notre cerveau regroupe en premier lieu les points les plus proches les uns des autres. Sur la Figure 6 (c), nous apercevons deux colonnes, alors qu'ils ne s'agit que de points.

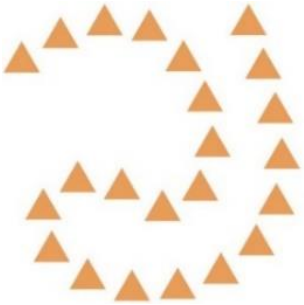
La loi de similitude : si la distance ne permet pas de regrouper les points, notre cerveau s'attache dans un second temps à repérer les plus similaires entre eux pour percevoir une forme. Dans la Figure 6 (d), notre cerveau va regrouper les objets grâce au critère de de la couleur, on observe alors une croix et quatre carrés.

La loi de destin commun : des parties en mouvement ayant la même trajectoire sont perçues comme faisant partie de la même forme. Dans la Figure 6 (e) notre cerveau regroupe entre elles les 3 lignes du haut et les 3 lignes du bas.

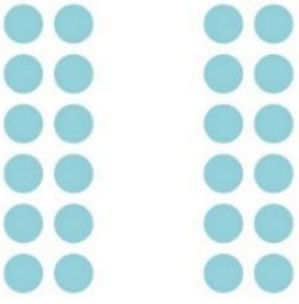
La loi de clôture : une forme fermée est plus facilement identifiée comme une figure (ou comme une forme) qu'une forme ouverte (cf. Figure 6 (f)).

Ces règles, permettant de relier des formes entre elles, vont être utilisées dans les différentes méthodes pour fusionner certains éléments du contenu du document, pour ainsi extraire la mise en page.

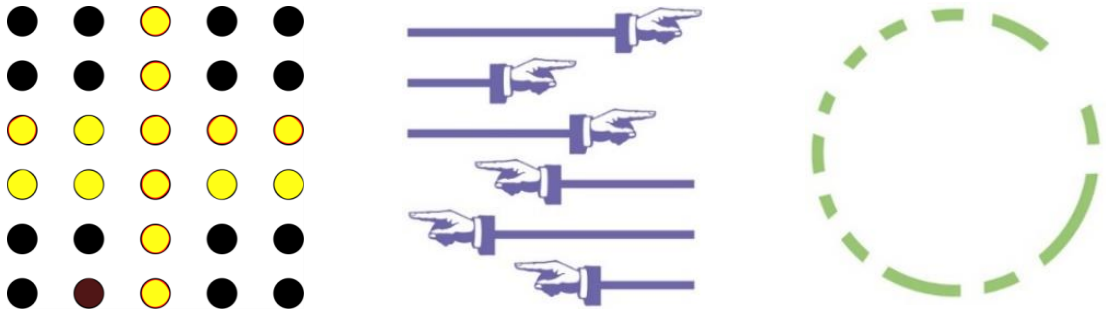
(a)



(b)



(c)



(d) (e) (f)

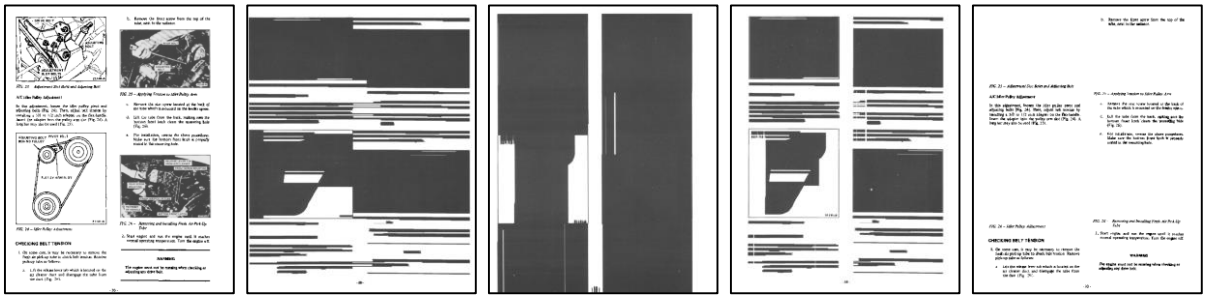
Figure 6 - Illustrations des lois de la Gestalt extraite de [3]. (a) Loi de bonne forme (vase de Rubin). (b) Loi de continuité. (c) Loi de proximité. (d) Loi de similitude. (e) Loi de destin commun. (f) Loi de clôture.

➢ **Approche par Run Length Smoothing Algorithm (RLSA)**

Cette méthode se base sur la loi de proximité de la Gestalt. La méthode *Run Length Smoothing Algorithm (RLSA)* ne s'applique que sur des images binaires de documents. Elle repose sur l'hypothèse que, dans le texte, une ligne horizontale est constituée de caractères qui, par proximité, constituent des mots avant de constituer une ligne. De manière à mettre en évidence ces entités, la méthode proposée par Wahl *et al.* en 1982 [WaWC82] pour segmenter le document en régions, consiste à noircir les espaces blancs, entre deux pixels noirs, dont la distance est inférieure à un seuil donné verticalement ou horizontalement selon que l'on veut mettre en évidence des paragraphes ou des lignes de texte. Il suffit ensuite d'opérer une intersection entre les deux images noircies pour obtenir les différentes zones segmentées (cf. Figure 7).

En ajustant le seuil, paramètre du RLSA qui sert à fusionner les composantes connexes, on peut obtenir différentes entités comme les mots, les lignes, *etc.* L'inconvénient de cette méthode est que les seuils sont définis empiriquement et cela de manière globale pour tout le contenu d'un document. Les valeurs optimales en fait dépendent du contenu de ce dernier. Il faudrait donc pouvoir les fixer de manière adaptative et locale. Par ailleurs, cette méthode ne peut donner qu'une segmentation binaire (par exemple texte et en non-texte). De plus, cette méthode n'est pas adaptée à une image d'entrée dont le contenu est incliné.

[3] https://www.nundesign.fr/pedagogie/fondamentaux-graphiques/gestalt-la-theorie-de-la-forme.



(a) (b) (c) (d) (e)

Figure 7 - Illustrations des étapes de la méthode RLSA sur un document (Images extraites de [WaWC82]*). (a) Image originale. (b) Méthode RLSA appliquée dans la direction horizontale. (c) Méthode RLSA appliquée dans la direction verticale. (d) Résultat de la segmentation en blocs. (e) Résultat des blocs considérés comme du texte.*

Cette méthode est encore très utilisée de nos jours. Grzejszczak *et al.* [GrRB12] et Hamrouni *et al.* [HaCV14] l'ont employée dans le processus de séparation entre le manuscrit et l'imprimé dans des documents administratifs.

➢ **Approches multi-résolutions**

Bloomberg [Bloo91] a proposé une approche multirésolution pour segmenter une image de document. Les approches multirésolutions sont des approches pyramidales qui traitent le document à différentes résolutions pouvant être obtenues de différentes manières. Pour la segmentation, des opérations morphologiques classiques et une nouvelle opération : l'ouverture généralisée (qui est efficace pour l'extraction de formes et de textures à une échelle donnée), sont utilisées. Les opérations multirésolution consistent en une généralisation des opérations morphologiques (appelées filtres par ordre de rang), suivies par un sous-échantillonnage. Cette approche sépare grossièrement l'image du document en texte et non-texte.

Bloomberg est également à l'origine d'une librairie open source « Leptonica »[4] qui a servi dans de nombreux articles de référence ainsi que dans des études comparatives [BASB10] et [ZENM13].

➢ **Approches par analyse du fond**

Ces méthodes se basent sur les lois de clôture et de proximité de la Gestalt. Un document est généralement constitué d'un ensemble de données, textes, graphiques ou images qui, pour une lecture aisée, sont séparés par des zones blanches servant de séparateur, facilitant la perception visuelle du document. Zones blanches et zones occupées sont donc duales. Ainsi, pour segmenter un document, il est également possible d'étudier les espaces blancs comme l'ont

[4] http://www.leptonica.com/.

fait Phillips et Zhou [PhZh91]. Les auteurs partent du principe qu'il n'est pas nécessaire de caractériser les espaces que l'on recherche car ils peuvent être isolés. Leur méthode détecte donc les grandes plages d'espaces blancs pour les regrouper.

Antonacopoulos [Anto98] a amélioré cette technique en ne cherchant plus des espaces blancs rectangulaires mais des espaces polygonaux, ce qui lui a permis de traiter les documents inclinés (cf. Figure 8). Cette technique, permet d'obtenir des résultats pertinents si le contenu du document est clairement délimité, mais a toutefois l'inconvénient d'être sensible au bruit produit notamment lors de l'étape de binarisation.

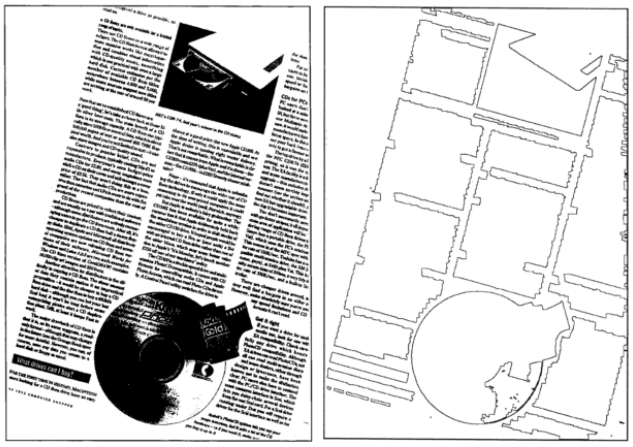


Figure 8 - Résultat de segmentation de la méthode de segmentation d'Antonacopoulos extraite de [Anto98].

➢ **Approches par analyse des composantes connexes**

Ces méthodes se basent sur la loi de similarité de la Gestalt. Dans cette section nous avons choisi de présenter deux méthodes qui utilisent, pour segmenter le document, la position relative des composantes connexes les unes par rapport aux autres et leur taille. La plupart des méthodes qui se fondent sur l'analyse d'images binaires utilisent les composantes connexes et pourraient être mentionnées ici.

Zirari *et al.* [ZENM13] proposent une méthode pour segmenter un document en texte et en non-texte. Cette méthode, insensible à une faible inclinaison du document, se base sur l'analyse des composantes connexes. Elle part du principe que les zones de texte sont souvent caractérisées par un alignement des caractères de tailles très similaires. C'est pourquoi l'approche commence par calculer un histogramme des fréquences de taille des composantes connexes, où l'on ne garde que les composantes correspondant aux modes de l'histogramme, et l'on termine l'opération en éliminant le bruit restant par la notion d'alignement des composantes. Cet alignement est déterminé par le chevauchement vertical entre les

composantes, en fonction d'un seuil donné par l'utilisateur, ce qui permet d'accepter une certaine inclinaison des documents.

Vauthier et Belaïd [VaBe12] cherchent à segmenter des images de documents en extrayant la couche textuelle grâce à l'analyse de la sortie d'un OCR (*Optical Character Recognition*, reconnaissance optique de caractères en français). Ils utilisent leur propre système pour regrouper les mots en lignes en utilisant des paramètres fixés qui ont été déduits empiriquement. Une étape de vérification orthographique est effectuée pour supprimer le bruit et l'écriture manuscrite qui auraient pu être à tort reconnus par l'OCR. Ces zones ainsi identifiées sont retirées de l'image. Les zones restantes sont regroupées par composantes connexes en fonction de leurs distances relatives et de leur taille (tout en éliminant les composantes de taille inférieure à 4×4 pour enlever le bruit) pour former des régions qui seront par la suite classifiées en tableau, signature, tampon et logo comme nous le verrons dans la section 1.3.4.

Ces deux méthodes ont toutes les deux pour inconvénient de ne s'appliquer que sur des images binaires : la première sert à identifier la couche textuelle pour aider l'OCR tandis que la seconde se sert de l'OCR pour trouver la couche textuelle.

La méthode de Melinda *et al.* [MeGB17] segmente les documents. Elle est fondée sur la hauteur des éléments et gaussiennes avec des mises en pages non Manhattan. En partant du principe que le document est composé en majorité de corps de texte (qui est généralement de taille fixe), de titre et de graphique, les auteurs ont cherché à prendre en compte les différences entre les éléments. La méthode est composée de 5 étapes : binarisation, recherche des composantes connexes, création de l'histogramme des hauteurs de lignes, recherche des n gaussiennes de l'histogramme (n étant un paramètre de la méthode (entre 2 et 5)), classification de ces gaussiennes et fusions des composantes connexes. Les plus petits éléments sont la ponctuation et le bruit, et la classe comportant le plus d'éléments est considérée comme le corps du texte. Ces éléments sont ensuite fusionnés en considérant leurs plus proches voisins pour donner des blocs rectangulaires. Les éléments graphiques sont les éléments de la dernière gaussienne (celle comportant les plus grands éléments). Les limites de la méthode sont qu'elle ne segmente que des blocs et les éléments en textes inversés.

1.2.3 Méthodes hybrides

Les méthodes hybrides sont des méthodes qui combinent l'approche ascendante et l'approche descendante. Nous présenterons dans cette catégorie une méthode utilisant la forme

et le fond du document pour segmenter le document, les méthodes fondées sur le diagramme de Voronoï et une méthode souvent utilisée comme référence : la méthode des *tab-stops*.

➢ **Méthode utilisant le fond et la forme**

Une autre façon, actuellement en vogue pour classer les méthodes de segmentation, est de considérer les méthodes qui partent de la forme et celles qui partent du fond. Cela suppose d'avoir au préalable binarisé l'image : on considérera les pixels noirs comme faisant partie de la forme et les pixels blancs comme du fond.

La méthode de Chen *et al.* [ChYL13] propose d'utiliser à la fois les grands espaces blancs trouvés en analysant le document dans sa globalité mais également les composantes connexes de formes pour segmenter les documents. Elle se compose de trois parties : l'extraction, le filtrage et le regroupement des rectangles de fond (espaces blancs). En premier lieu, les composantes connexes de la forme sont extraites et reliées selon leur relation d'adjacence horizontale. Les rectangles composés d'espaces blancs sont extraits de l'espace entre les composantes connexes adjacentes horizontalement, et sont progressivement filtrés en fonction de la comparaison de la largeur du rectangle et de la relation d'adjacence avec les lignes de texte. Les rectangles restants sont regroupés en séparateurs et les séparateurs non viables sont filtrés de manière heuristique. Les grandes composantes connexes sont ensuite fusionnées avec des lignes de texte. Enfin, les blocs de texte sont formés en groupant les lignes de texte et en les ordonnant.

➢ **Approches basées sur le diagramme de Voronoï**

Le diagramme de Voronoï est un outil qui permet un découpage du plan (pavage) qui a pour particularité de séparer chacun des points ou ensemble de points (éléments dits germes) dans une cellule pour que tous les éléments d'une cellule soient plus proches de leur germe que des autres.

Kise *et al.* [KiSI98] se basent sur le diagramme de Voronoï pour segmenter les documents. L'approche proposée par les auteurs commence par extraire les frontières (contours) des composantes connexes pour calculer le pavage de Voronoï. Les germes sur la Figure 9 sont les contours des composantes connexes. La fusion des cellules se base sur une étude statistique des distances entre les cellules. La densité de chaque cellule est calculée afin d'émettre des hypothèses sur les espaces inter-lettres, inter-mots, inter-lignes et de fusionner les cellules en conséquence, comme le montre la Figure 10.

Les avantages majeurs de cette méthode sont qu'elle permet de traiter des documents même à forte inclinaison et que son découpage est plus fin. Cela permet ainsi de traiter des documents avec des mises en page plus complexes.

Dans leurs travaux, Winter *et al.* [WiAS11] ont amélioré la reconnaissance de caractères en segmentant les documents avec cette technique.

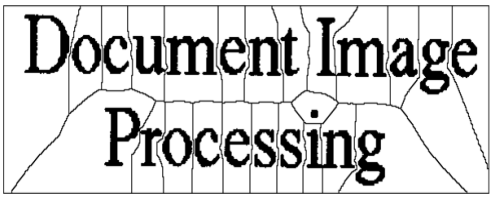
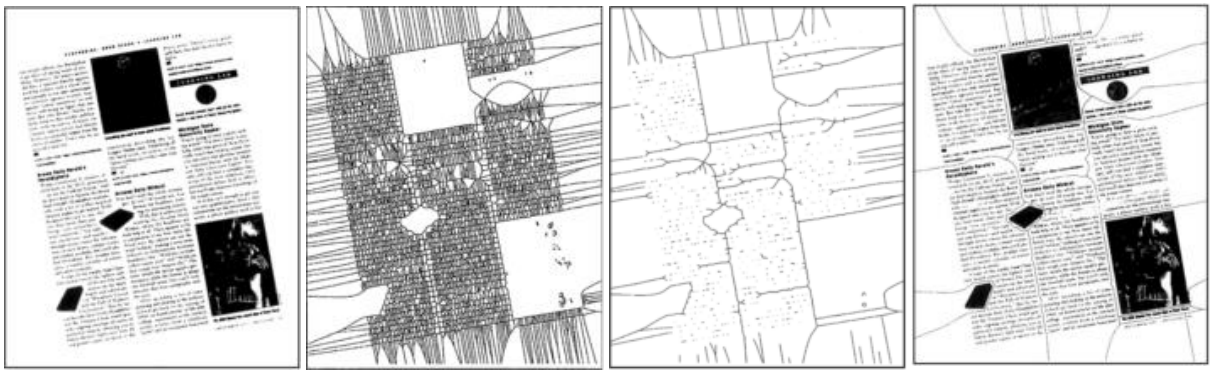


Figure 9 - Diagramme de Voronoï (image extraite de [KiSI98]*) construit sur « Document Image Processing » en choisissant pour germe les contours des composantes connexes de l'image binaire.*



(a) (b) (c) (d)

Figure 10 - Illustrations des étapes de la segmentation par le diagramme de Voronoï (images extraites de [KiSI98]*). (a) Document original scanné. (b) Diagramme de Voronoï construit sur l'image. (c) Diagramme après suppression des arêtes superflues. (d) Résultat sur l'image de la segmentation à base de diagramme de Voronoï.*

➢ Approches fondées sur les *tab-stops*

Les *tab-stops* sont les composantes qui commencent et terminent les lignes de texte. Les trouver permet, en calculant l'alignement, de trouver les lignes et ainsi de segmenter un document à ce niveau.

Smith [Smit09] présente une méthode de segmentation d'un document se basant sur les tab-stops. Il commence par une étape de prétraitement pour identifier des lignes ou des séparateurs à dominantes horizontales et verticales, et pour localiser les régions en demi-teinte ou en image dans le document. Ensuite, une analyse des composantes connexes est réalisée afin d'identifier les *tab-stops* en fonction de la taille et de la largeur de trait.

Les composantes de texte sont évaluées en tant que candidats comme tab-stop. Ces candidats sont regroupés en lignes verticales afin de trouver les positions des taquets de

tabulation qui sont alignés verticalement. L'étape finale consiste alors à trouver les paires de *tab-stops* qui forment les lignes, à les relier et à les ajuster de telle sorte que les *tab-stops* finissent dans la même abscisse. Les lignes de tabulation verticales marquent le début et la fin des zones de texte. Les lignes qui ont une taille de police et un espacement entre les lignes différentes sont regroupées en différents blocs. Ensuite, l'ordre de lecture de ces blocs est identifié (cf. Figure 11).

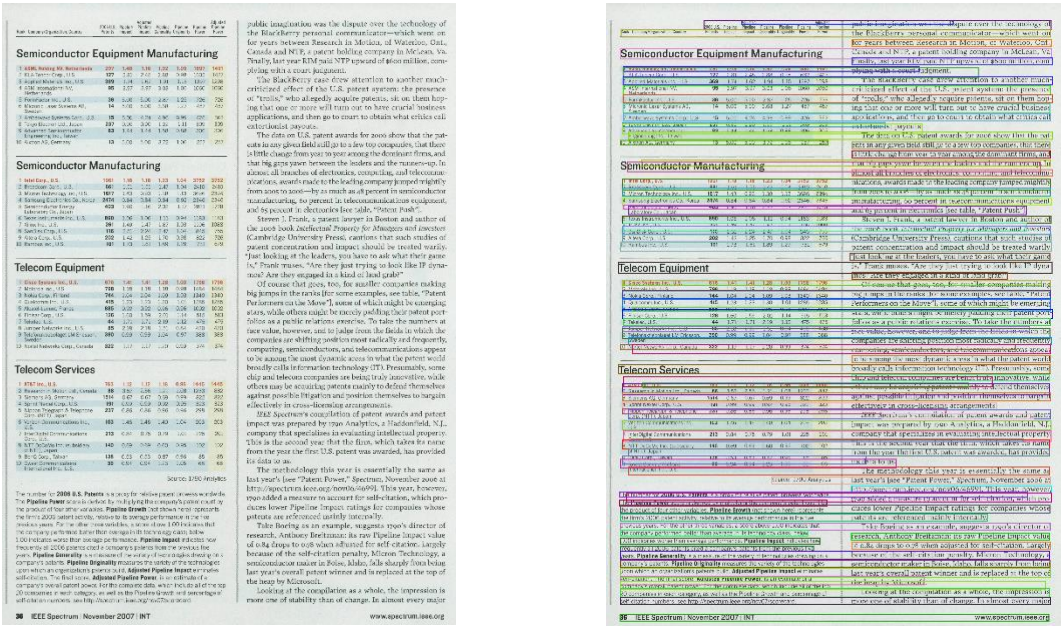
L'implémentation de cette méthode est disponible en open source en langage C++ dans la librairie Tesseract[5] ce qui la rend très populaire pour un grand nombre de chercheurs qui continuent de comparer leurs méthodes à elle.

(a) (b) (c) (d) (e) (f)

Figure 11 - Illustrations des étapes de la segmentation par les tab-stops (images extraites de [Smit09]*). (a) Composantes tab-stop candidates. (b) Lignes de textes liant les tab-stops. (c) Partition en colonnes et lignes tab-stop. (d) Colonnes. (e) Types de sous-colonnes. (e) Régions.*

Lors de la phase d'expérimentations préliminaires du projet SHADES, nous avons testé cette méthode sur différents types de documents. Nous avons pu observer qu'en présence de tableaux, la méthode tab-stop implémentée dans Tesseract ne parvenait pas à fournir des résultats pertinents sur la segmentation de certains documents, principalement dans des documents structurés en deux colonnes. Comme nous pouvons le voir sur la Figure 12, la plupart des lignes extraites (représentées par des cadres de couleurs aléatoirement choisies) par cette méthode commencent au début de la première colonne pour finir sur la deuxième. Cependant, dans la plupart des autres cas, cette implémentation offrait des résultats satisfaisants pour nos besoins.

[5] https://code.google.com/p/tesseract-ocr/.



(a) (b)

Figure 12 - Résultat de la segmentation de lignes de textes fondée sur les Tab-Stops (b) appliqué sur une image originale (a) dans Tesseract pour extraire les lignes (encadrer par des couleur aléatoires) en présence de tableaux.

1.2.4 Conclusion

Dans cette section nous avons mentionné différentes méthodes pour segmenter les documents en différentes régions regroupées selon trois idées : les méthodes descendantes, les méthodes ascendantes et les méthodes hybrides reprenant l'idée de partir localement des régions tout en gardant un point de vue plus global pour combiner les avantages des deux points de vue. Nous avons résumé dans le Tableau 1 les avantages et les inconvénients des différentes méthodes. Ces méthodes possèdent des avantages et des inconvénients propres mais elles ne sont pas assez généralisables et possèdent des inconvénients qui empêchent leur utilisation dans ce projet.

Nous avons vu dans un premier temps les méthodes de segmentation permettant de découper un document en différentes régions. Dans la section suivante, nous introduisons les méthodes d'extraction de la mise en page par classification, que ce soit en utilisant les zones précédemment extraites ou non.

Tableau 1 - Synthèse des principales méthodes de segmentation dans les images de documents.

Méthodes	***Avantages***	***Inconvénients***
X-Y cut	- Segmente en blocs	- Sensible à l'inclinaison du contenu des documents - Sensible aux bruits - Applicable sur des documents Manhattan
RLSA	- Sépare en texte et non texte - Temps de traitement assez court	- Nécessite une orientation horizontale du texte - Réglage des seuils de lissage
Multi-résolutions	- Extrait une couche textuelle	- Binarisation
Analyse du fond	- Accepte inclinaison	- Sensible aux bruits - Difficile à paramétrer
Analyse des CC	- Extrait une couche textuelle	- Binarisation - Difficile à paramétrer
Analyse du fond et de la forme	- Segmente en ligne de texte	- Binarisation
Tab-stop	- Segmente le texte en lignes et en blocs	- Perturbée par les tableaux
Voronoï	- Accepte les inclinaisons - Sépare en blocs	- Sensible au bruit

1.3 État de l'art sur les méthodes d'extraction de la mise en page par classification

Les méthodes de classification issues du domaine de l'apprentissage *(machine learning)* sont généralement employées pour catégoriser les différents contenus des documents en un ensemble de classes (comme les lettres, les factures, les articles, *etc.*). Le principe général de ces méthodes est de trier (ou de séparer) un ensemble de données pour les regrouper selon différentes classes/étiquettes en fonction de leurs caractéristiques, comme les caractéristiques de taille, de couleur, *etc.* La classification se fait le plus souvent à partir d'un apprentissage. Il existe trois types d'apprentissage :

- supervisé : la méthode de classification apprend un modèle sur un jeu de données dont on connaît déjà la classification (les données sont étiquetées) pour produire un modèle de prédiction ;
- non-supervisé : la classification se fait sur des données dont on ne connaît pas la classe finale. La méthode essaie donc de classifier selon des caractéristiques extraites, souvent sans *a priori* sur la structure sous-jacente (dans l'espace des caractéristiques) des classes d'intérêt recherchées ;
- semi-supervisé : la classification se fait sur un jeu de données dont une partie est étiquetée et l'autre non.

Comme dans tout problème de classification, on peut distinguer deux éléments principaux, la nature du classifieur et les caractéristiques utilisées. Lors de l'apprentissage, il est essentiel de séparer les données, sur lesquelles les algorithmes apprennent les valeurs des caractéristiques, des données sur lesquelles les tests sont effectués pour valider l'approche. Le surapprentissage correspond à un apprentissage parfait des données utilisées pour apprendre. Les résultats alors sont très bons si on redonne au classifieur une entrée identique à celles qu'il a apprises. Ils sont cependant très mauvais si l'entrée s'en éloigne même légèrement, il n'y a pas de pouvoir de généralisation.

Les approches par classification peuvent également être considérées pour extraire la mise en page d'un document sans utiliser de segmentation auparavant. C'est le cas des méthodes qui cherchent à étiqueter les pixels de l'image contrairement aux méthodes de classification par zones qui reposent sur une pré-étape de segmentation.

1.3.1 Algorithmes d'apprentissage

Il existe de nombreux algorithmes d'apprentissage dans la littérature. Dans cette partie, nous aborderons rapidement les différents algorithmes d'apprentissage utilisés dans la labélisation des pixels (cf. Section 1.3.3) et des régions (cf. Section 1.3.4) comme la méthode des plus proches voisins, les machines à vecteurs de support, les arbres de décision, les algorithmes de *boosting* et les réseaux de neurones.

- **K plus proches voisins (K-NN)**

L'algorithme des plus proches voisins (en anglais *K-Nearest Neighbors*) est une méthode d'apprentissage supervisé se basant sur l'idée que les éléments d'une même classe sont plus proches les uns des autres dans l'espace des caractéristiques que des autres classes. Elle se base sur le principe « dis-moi qui sont tes voisins, je te dirai qui tu es ». L'algorithme K-NN classe les entrées en fonction du jeu de données entier. Il va chercher pour chaque entrée ses K plus proches voisins. Une métrique de distance est ainsi choisie pour évaluer la distance entre les différents éléments comme la distance euclidienne, la distance de Manhattan, la distance de Minkowski, *etc.* Le choix de cette distance est important et peut modifier les résultats obtenus. Les plus proches voisins trouvés permettent ainsi de classer la nouvelle entrée en fonction des labels trouvés. Pour utiliser cet algorithme, deux paramètres doivent être considérés : le nombre de voisins à considérer et le mode de prédiction. Ainsi sur la Figure 13, la donnée représentée par le disque vert serait attribué à la classe triangle rouge si K = 3 et carré bleu si K = 5 si l'on considère pour la prise de décision, le nombre majoritaire d'éléments parmi les k voisins.

Comme dans tous les classifieurs, les données (dans notre cas les données sont composées par les caractéristiques extraites des zones) sont représentées par un vecteur de caractéristiques.

Ce classifieur est simple à comprendre et à adapter lorsque les données peuvent être correctement rassemblées en label dans un espace de caractéristiques définies. L'inconvénient de la méthode est que tout le jeu de données d'apprentissage doit être gardé en mémoire pour effectuer une prédiction.

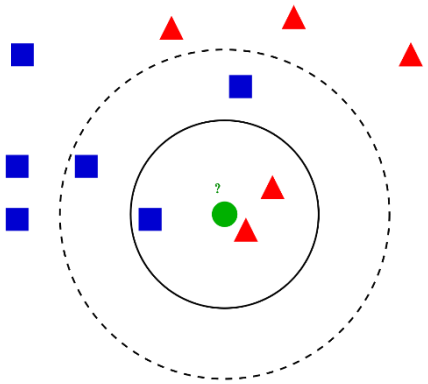


Figure 13 - Exemple d'attribution d'une classe à un nouvel élément (disque vert) par une classification de KNN sur un ensemble à 2 classes (représentées par des carrés bleus et des triangles rouges).

➢ Machines à Vecteurs de Support (SVM)

Les Machines à Vecteurs de Support (en anglais *Support Vector Machine*) sont un ensemble de méthodes d'apprentissage supervisé. Ces méthodes cherchent à créer des frontières dans l'espace des caractéristiques pour séparer l'ensemble des données en différentes classes. Ils évoluent autour de la notion de marge. Dans le cas de données facilement séparables les frontières sont des hyperplans mais dans les autres cas, des noyaux sont utilisés pour améliorer la séparabilité des données. La Figure 14 montre un jeu de données contenant 2 classes : les points blancs et les points noirs et 3 hyperplans pour les séparer. Nous observons que l'hyperplan H3 ne permet pas de séparer les données. En effet, il sépare seulement un élément blanc du reste du jeu de données. Les hyperplans H1 et H2 permettent de séparer correctement les données, mais nous remarquons que H1 le fait avec des marges plus réduites que H2. L'hyperplan maximisant les marges que le SVM va construire dans ce cas-ci est H2.

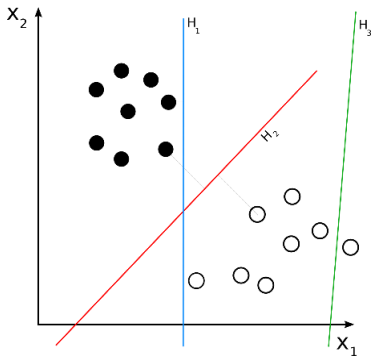


Figure 14 - Exemple de 3 hyperplans séparateurs pour séparer un ensemble de 2 classes (points blancs et noirs). H1 sépare les classes avec une petite marge, H2 avec une marge optimale et H3 ne sépare pas efficacement les données par rapport aux 2 classes.

Un avantage des Machines à Vecteurs de Support est leur capacité à traiter des problèmes non linéaires grâce à l'utilisation de noyaux. Ces approches sont efficaces dans les espaces de grandes dimensions. Comme pour la méthode précédente, elles ne sont pas efficaces dans le cas de classes bruitées.

➢ Arbres de décision

Les arbres de décision permettent l'affectation d'un label à une donnée à travers une procédure de décision hiérarchique. Le procédé de classification peut être décrit au moyen d'un arbre, dans lequel au moins un nœud terminal est associé à chaque classe et chaque branche à la réponse d'une caractéristique. La Figure 15 illustre un exemple d'arbre de décision permettant de prédire si des joueurs vont jouer ou non en considérant 14 observations météorologiques en fonction des caractéristiques « ensoleillement », « humidité » et « vent ». Dans le cadre d'un apprentissage supervisé, l'arbre de décision est déterminé automatiquement.

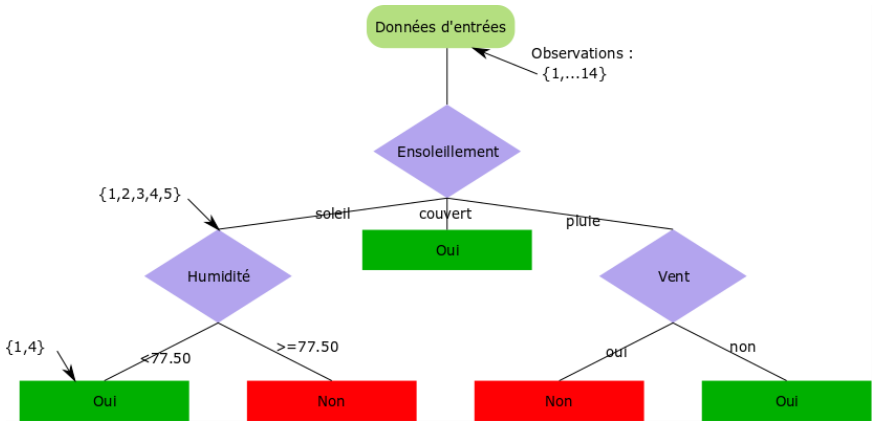


Figure 15 - Exemple d'arbre de décision de données météorologiques (14 observations) permettant de prédire la variable « jouer » (oui en vert et non en rouge) en fonction des caractéristiques « ensoleillement », « humidité » et « vent ».

Ces méthodes sont simples à comprendre et à interpréter. Elles ne demandent pas de normalisation entre les différentes caractéristiques car celles-ci sont traitées indépendamment. Elles sont également performantes sur un grand jeu de données. Une fois l'arbre construit, il est facile de prédire le résultat grâce à celui-ci. Les inconvénients de ces méthodes tiennent dans la difficulté d'optimiser l'ordre des caractéristiques permettant de créer les nœuds. Un arbre de décision non optimal peut correspondre à un surapprentissage.

Forêts aléatoires

Les forêts aléatoires (ou *random forest* en anglais) sont des méthodes d'apprentissage utilisant plusieurs arbres de décision pour classifier les données. Les arbres de décision sont construits à partir de différents sous-échantillons de l'ensemble d'apprentissage et des caractéristiques. La décision d'attribution d'une entrée à une classe est effectuée par un vote majoritaire. Ainsi l'entrée est classifiée par chaque arbre de décision qui prédit une classe et la classe majoritaire sera attribuée à l'entrée.

Les avantages de ces méthodes sont généralement de meilleurs résultats qu'avec les arbres de décision, moins de surapprentissage mais les inconvénients sont l'apprentissage lent et elles sont moins lisibles que les arbres de décisions.

- **Algorithme de Boosting**

Les algorithmes de Boosting apparus dans les années 80 regroupent des méthodes capables de prendre des décisions précises à partir de règles de décision dites "faibles". Ces règles ont généralement un pourcentage de réussite d'au moins 50% si la distribution des classes dans le jeu de données est équilibrée. L'algorithme en plusieurs itérations cherche des règles pour les données mal classées.

Les avantages des algorithmes de Boosting résident dans leur simplicité et leur facilité d'utilisation. L'erreur reste stable même après de nombreuses itérations, mais ils ne sont pas adaptés à une utilisation sur un petit jeu de données.

- **Réseaux de neurones**

Les réseaux de neurones (en anglais *artificial neural network*) s'inspirent du cerveau humain pour classifier les informations. Ils sont organisés selon des neurones et des synapses. Lors de la phase d'apprentissage, le système apprend sur tous les exemples de l'ensemble d'apprentissage de données jusqu'à ce que les données d'apprentissage ne modifient plus le système. On obtient ainsi une convergence.

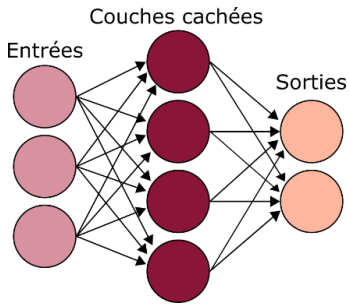


Figure 16 - Illustration d'un réseau de neurones à 1 couche avec 3 entrées et 2 sorties.

Deep learning

Le *deep learning* repose sur un réseau de neurones qui permet d'apprendre à partir d'informations brutes en grandes quantités. Cette méthode d'apprentissage est très récente. Elle passe par un apprentissage à plusieurs niveaux de détails ou de représentations des données. À travers les différentes couches, on passe de paramètres de bas niveau à des paramètres de plus haut niveau. Cette technique commence à être utilisée dans l'analyse de la mise en page, comme l'ont fait Chan *et al.* [CSHI17] pour traiter des documents anciens.

D'un point de vue général, le projet SHADES a pour objectif de traiter tout type de documents, l'apprentissage porterait sur des données hétéroclites. Cela peut être un inconvénient dans le cadre d'un apprentissage. De plus le *deep learning* repose sur une boîte noire dans le sens où l'on ne sait pas quelles caractéristiques sont utilisées.

1.3.2 Caractéristiques utilisées

Les caractéristiques souvent utilisées pour la classification des zones d'un document sont des caractéristiques de texture, des caractéristiques de couleur ou des caractéristiques géométriques. L'analyse de la texture a été très étudiée en analyse d'images car elle est utilisée entre autres pour segmenter les images naturelles.

➢ Caractéristiques de texture

On peut définir la texture comme la répétition d'un motif dans différentes directions de l'espace (cf. Figure 17). Il existe un très large panel de caractéristiques car la texture est utilisée dans de nombreux domaines comme la segmentation ou la compression des images naturelles. Des chercheurs se sont intéressés à essayer de décrire « visuellement » la texture en utilisant le vocabulaire suivant : grossière, fine, lisse, tachetée, granuleuse, marbrée, régulière ou irrégulière. Nous allons présenter dans cette section les caractéristiques liées à la texture utilisées dans l'extraction de la mise en page comme celles utilisant les *run length*, les matrices

de co-occurence, *Tamura texture feature histogramm* ou l'auto-corrélation. Il en existe bien d'autres, par exemple, celles issues des transformées de Gabors pour n'en citer qu'une.

Figure 17 - Exemple de textures des bases Brodatz[6] (haut) et VisTex[7] (bas).

Run length : Les *run length* est une méthode fondée sur la redondance. Le principe est de considérer les suites d'informations. Ainsi dans les documents binaires, les *run length* donnent les longueurs de segments calculés dans une direction donnée (généralement sur quatre directions : horizontale, verticale, diagonale-gauche et diagonale-droite). Ils sont utilisés pour reconnaître des formes dans les zones. Il y a différentes façons d'utiliser ces longueurs des segments. L'une de ces façons est de calculer leur longueur moyenne et leur variance.

Matrice de co-occurrence : Les matrices de co-occurence permettent de caractériser la périodicité et la direction des textures. Elles peuvent être calculées sur une image à niveaux de gris mais également sur une image binaire. On parle alors en anglais de *bi-level co-occurence*. Il n'y a que quatre paires de pixels possibles : noir-noir, noir-blanc, blanc-noir et blanc-blanc.

Tamura texture feature histogramm : Les caractéristiques de Tamura se basent sur six caractéristiques visuelles : grain (*coarseness*), contraste (*contrast*), direction (direction), présence de lignes (*line-likeness*), régularité (*regularity*), rugosité (*roughness*). Le grain est une mesure de la granularité de l'image. Le contraste mesure la variation des niveaux de gris dans l'image. Le paramètre de direction mesure l'orientation principale dans l'image et non les différentes directions qui peuvent éventuellement exister dans l'image. Ce paramètre détecte si l'image comprend des lignes ou non.

[6] http://www.ux.uis.no/~tranden/brodatz.html.
[7] http://vismod.media.mit.edu/vismod/imagery/VisionTexture/vistex.html.

Autocorrélation : La fonction d'autocorrélation est définie comme étant la corrélation croisée d'un signal avec lui-même, et représente une mesure de similitude entre les deux signaux. Une fois appliquée à une image en niveaux de gris, on produit une matrice symétrique centrale, qui donne une idée du degré de régularité de la texture. La fonction d'autocorrélation est efficace pour trouver les motifs répétitifs. Dans le cas des documents, cette caractéristique est adaptée puisque les textures textuelles ont une orientation prononcée qui diffère fortement de celles des illustrations.

La méthode consiste à subdiviser l'image originale I en blocs carrés de taille paire n. La définition formelle de l'autocorrélation d'un bloc est définie par :

$$C(k,l) = \sum_{y=\max(0,l)}^{n-1+\min(0,l)} \sum_{x=\max(0,k)}^{n-1+\min(0,k)} I(x,y) \cdot I(x-k, y-l)$$

ou l et k sont définies sur $\left[-\frac{n}{2}, \frac{n}{2}\right]$.

En général, la matrice d'autocorrélation est codée avec un histogramme de la direction, une représentation polaire, dans laquelle chaque direction est déterminée par un angle [0°, 360°] et la valeur de chaque élément de l'histogramme est donnée par la somme des niveaux des pixels le long de cette direction.

La Figure 18 montre un résumé visuel du processus d'extraction de caractéristiques et quelques exemples de résultats de Coppi *et al.* [CoGC14].

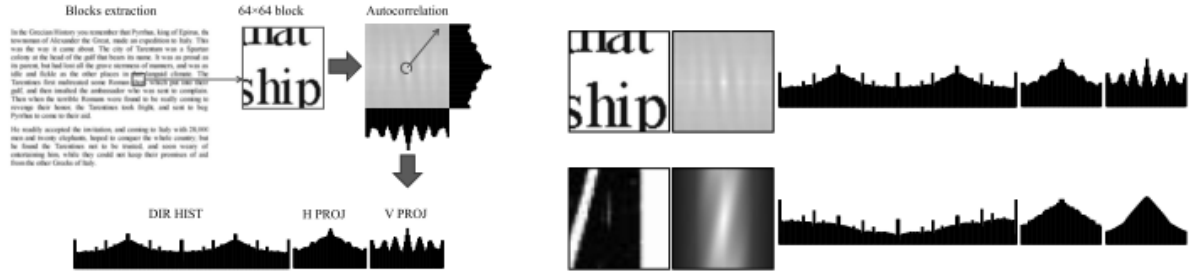


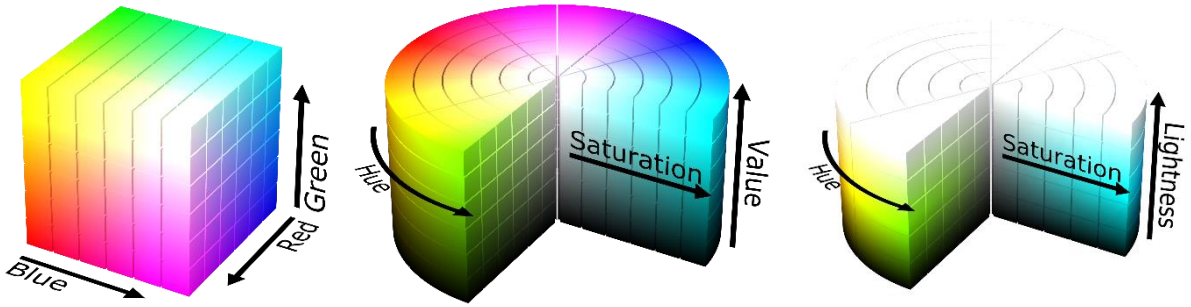
Figure 18 - Illustration de l'auto-corrélation extraite de [CoGC14] *(a) Représentation du calcul des vecteurs de caractéristiques calculés à partir d'un bloc sur une image. (b) Illustration des vecteurs de caractéristiques calculés sur du texte (en haut) et sur une illustration (en bas).*

➢ Caractéristiques de couleur

La couleur est une indication importante pour identifier le contenu d'un document, son uniformité peut par exemple permettre la différentiation des éléments recherchés. Un autre moyen d'analyser la couleur est de changer l'espace de représentation. En effet, classiquement nous utilisons l'espace de représentation RGB (cf. Figure 19 (a)) indiquant les niveaux de bleu,

vert et rouge dans chaque canal radiométrique mais ce n'est pas le seul espace de représentation. Il existe par exemple l'espace de Teinte Saturation Valeur *(Hue Saturation Value (HSV))* (cf. Figure 19 (b)) ou l'espace de Teinte Saturation Luminosité *(Hue Saturation Lightness (HSL))* (cf. Figure 19 (c)). L'espace HSL est une représentation des couleurs selon trois paramètres :

- la teinte correspondant à un angle compris entre 0° et 360° (0° correspondant au rouge, 120° au vert et 240° au bleu) ;
- la saturation correspondant à l'intensité de la couleur elle est généralement comprise entre 0 et 1 ;
- et la luminosité, elle aussi comprise entre 0 et 1 (0 correspondant au noir et 1 au blanc).



(a) (b) (c)

Figure 19 - Représentations de la couleur dans différents espaces, (a) dans l'espace RGB, (b) dans l'espace HSV et (c) dans l'espace HSL.

Dans l'espace HSL, de nombreuses caractéristiques peuvent être calculées. Baird *et al.* [BMAC07] ont étudié plus de soixante caractéristiques toutes extraites du canal de luminosité (en ignorant les canaux de teinte et de saturation) pour extraire la mise en page. Les auteurs détaillent l'utilité de ces caractéristiques mais n'en sélectionnent que vingt-six pour leur base de test. Celles-ci sont calculées sur chaque pixel en prenant en compte un voisinage défini en considérant la moyenne des valeurs ou par différences.

➢ Caractéristiques géométriques

Étude des composantes connexes : Les composantes connexes sont extraites à partir d'images binaires. On peut les récupérer selon la relation d'adjacence de 4-connexité ou celle de 8-connexité. On les utilise ensuite pour calculer des informations quantitatives comme leurs taille, inertie, densité, l'histogramme de distance entre une composante connexe et son plus proche voisin.

Caractéristiques statistiques : De nombreuses caractéristiques statistiques peuvent être extraites d'une région. La densité des pixels noirs, la hauteur de la région par exemple permettent de différentier les différentes couches du document.

1.3.3 Méthodes de classification fondée pixels

Les méthodes de classification sont applicables quelle que soit l'image du document à étudier. Les méthodes présentées ci-dessous ne requièrent pas de première étape de segmentation. Elles utilisent un algorithme pour traiter une page de document qui classe les pixels un par un afin d'obtenir en sortie une image labélisée, puis une étape de reconstruction par régularisation permet de fournir une image segmentée.

La classification par pixel a l'avantage de s'abstraire de toute forme *a priori*, contrairement à la classification par zones qui sous-entend le plus souvent d'avoir des formes prédéfinies pour les structures contenues dans le document. Mais ces méthodes ont malheureusement souvent le problème de positionner plusieurs classes dans un même voisinage conduisant à un effet *poivre et sel*. Ce constat suggère la nécessité d'un post-traitement qui force l'uniformité locale sans imposer de restreindre la forme des régions.

Ces méthodes sont constituées des étapes suivantes : chaque pixel est caractérisé suivant différentes propriétés concernant la texture, la couleur, le voisinage, *etc.* L'apprentissage se fait sur un grand nombre de documents permettant ainsi de construire un modèle prédictif. La mise en page de l'image est identifiée par classification des pixels (cf. Figure 20). Par ailleurs, il est à noter que le classement par pixels évite la limitation arbitraire de la forme d'un média, par exemple constitué de blocs de formes rectangulaires.

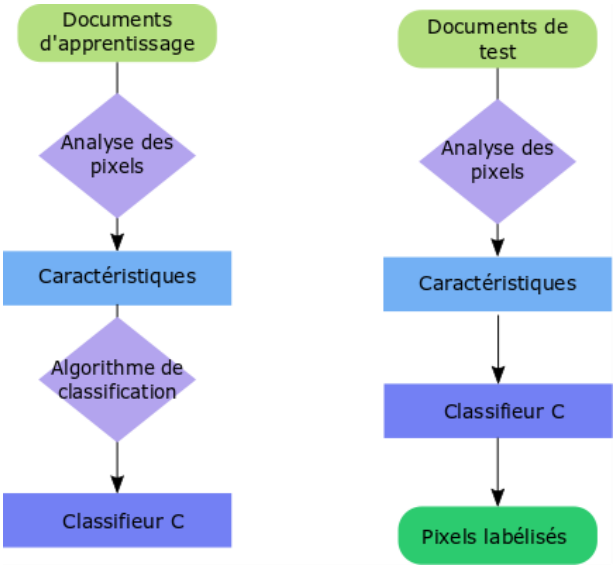


Figure 20 - Illustration de la chaîne de traitements des méthodes de classification des pixels.

Les méthodes présentées seront regroupées en fonction du type de caractéristiques qu'elles utilisent relevant de la colorimétrie, de la forme ou de la statistique.

Baird et *al.* [BMAC07] ont développé des techniques fondées pixels pour analyser le contenu d'un document en considérant les caractéristiques de luminance dans l'espace HSL. Les auteurs présentent l'impact des différents algorithmes de classification et des différentes caractéristiques sur le temps d'exécution et la précision. Ils emploient comme algorithme d'apprentissage de référence la méthode des K-NN qu'ils considèrent comme étant rapide et donnant une précision supérieure aux autres méthodes. Ces méthodes permettent de classifier des documents très complexes avec une mise en page non rectiligne. Malheureusement, ces méthodes ne sont pas pertinentes pour distinguer l'écriture manuscrite de l'écriture imprimée. Par ailleurs, elles confondent souvent les bords des images avec du texte manuscrit.

Cohen *et al.* [CAKE13] présentent une méthode pour segmenter les images de documents historiques. Ils commencent par séparer le texte du non-texte avec une binarisation puis classent le non-texte en trois classes : bruit, image et fond en utilisant des caractéristiques de forme et des caractéristiques de couleur. Pour obtenir un temps de calcul plus rapide, cette technique utilise les superpixels qui sont des unités élémentaires homogènes à partir desquels les caractéristiques locales de l'image peuvent être calculées. Ces zones de superpixels [ASSL12] se sont révélées de plus en plus utiles pour des tâches de vision par ordinateur, car elles réduisent grandement la complexité des étapes ultérieures de traitement d'images.

Vieux et Domenger [ViDo12] présentent une méthode pour classifier les pixels d'une image d'un document ancien en trois catégories : texte, image et fond. La méthode démarre par une étape d'apprentissage. Chaque image participant à l'apprentissage est traitée selon la démarche décrite ci-dessous : un vecteur de description contenant le résultat d'une banque de filtres de Gabor (avec 5 fréquences et 6 orientations) et un label (texte, image...) sont associés à chaque pixel de l'image traitée. Ces données sont ensuite traitées par un algorithme de classification non supervisé : K-means à 2 classes. Si dans une classe donnée par cet algorithme, il n'y a pas de label prédominant, c'est-à-dire qu'aucun label ne ressort significativement selon un seuil choisi par l'utilisateur, une nouvelle classification est effectuée (cf. Figure 21). Si une étiquette est très répandue dans le cluster, le centre de gravité du cluster, le poids du cluster (nombre de pixels) et la distribution d'étiquettes dans le modèle sont enregistrés. Un seuil maximum de récursivité est fixé pour éviter de multiples scissions au sein d'un ensemble difficilement séparable. Chaque image de l'ensemble de la base est traitée. Les centres de gravité associés, les poids et les étiquettes des distributions de fréquence sont sauvegardés. L'étape suivante fait appel à la méthode des k plus proches voisins (KNN). La méthode KNN utilise

alors le vecteur de description associé à chaque pixel comparé au vecteur associé aux centres de gravité enregistrés. Cette méthode peut traiter des documents dont les mises en page sont fortement complexes.

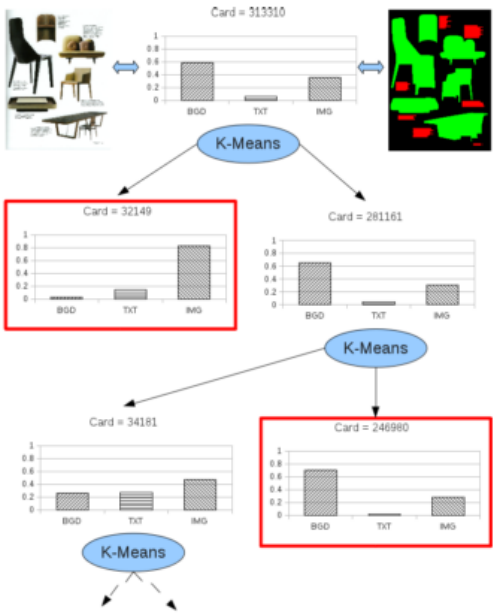


Figure 21: Illustration de la méthode d'apprentissage présentée dans Vieux et al. [ViDo12]

Cote et Branzan Albu [CoBr14] se sont intéressés à une méthode de classification des pixels pour classifier le contenu des documents administratifs en quatre classes : texte, image, graphique et fond. Pour cela, les auteurs utilisent les réponses des filtres de Leung-Malik et l'algorithme SVM pour la classification.

Caponetti *et al.* [CaCa08] proposent une méthode pour classer chaque pixel selon trois classes : texte, graphique et fond. Leur approche est fondée sur les réseaux de neurones flous et sur l'analyse d'un ensemble de caractéristiques extraites de l'image, disponibles à différents niveaux de résolution. Une segmentation initiale est obtenue en classant les pixels en régions cohérentes qui sont successivement affinées par l'analyse de leur forme. Les résultats obtenus montrent que cette méthode est robuste au bruit et à l'inclinaison des documents.

➢ Bilan et discussions

Nous avons présenté précédemment différentes méthodes de classification fondées pixels. Nous avons résumé dans le Tableau 2 les différences entre ces méthodes impliquant différents types de caractéristiques et leur nombre, le type de classifieurs et le nombre de couches qu'elles peuvent extraire. Dans les approches cherchant à classifier les pixels peu de couches sont extraites.

Tableau 2 - Synthèse des principales méthodes de classification des pixels dans les images de documents.

Méthodes	***Type de caractéristiques***	***Nombre de caractéristiques***	***Type de classifieurs***	***Couches extraites***
[BMAC07]	Couleur	26	KNN	- imprimé - manuscrit - photographie - fond
[CAKE13]	Forme et couleur	12	KNN	- texte - bruit - image - fond
[ViDo12]	Formes	30	SVM	- texte - image - fond
[CoBr14]	Texture	48	SVM	- texte - image - graphique - fond
[CaCa08]	Statistique	-	réseaux de neurones	- texte - graphique - fond

L'avantage majeur des méthodes de classification basées pixels est qu'elles nécessitent peu d'*a priori* sur la mise en forme du document à traiter, ce qui permet de traiter une plus large gamme de documents. Cependant elles ont comme inconvénient d'être généralement moins précises que les méthodes qui utilisent les propriétés des documents. Un autre inconvénient réside dans le fait qu'elles ont besoin d'un grand nombre d'images pour la phase d'apprentissage.

Les méthodes de classification fondées pixels acceptent généralement des images de documents dont le contenu est fortement incliné car elles ne font bien souvent pas d'*a priori* sur la forme que doit avoir le contenu. La classification par pixels est extrêmement liée au comportement des pixels qui peuvent présenter de nombreuses instabilités entre différentes instances d'un document hybride et donc poser problème dans le cadre du projet SHADES qui recherche une stabilité dans les résultats sortant d'instances d'un même document hybride. Ces méthodes ont généralement moins d'*a priori* sur le contenu que les autres méthodes que nous allons voir par la suite. Nous entendons par là qu'elles ne supposent pas que le document doit contenir tel ou tel média.

1.3.4 Méthodes de classification fondée sur les régions

Dans cette partie, il est supposé que les documents ont déjà été segmentés en régions (cf. Section 1.2). L'objectif des méthodes de classification par région est alors d'attribuer une étiquette / catégorie à chacun de ces blocs. Les labels de base dans l'analyse d'une image de

document sont : « texte » et « fond ». Les autres labels possibles sont « image », « logo », « graphique » et peuvent être encore plus spécifiques comme « bruit », « lettrines », « tampons », « signatures », *etc.*

Wang *et al.* [WaPH06] présentent une technique pour classifier les blocs d'un document. Les auteurs les classent en neuf catégories : texte de taille inférieure ou égale à 18 points, texte de taille supérieure à 18 points, formule mathématique, tableau, *halftone*, carte / dessin, trait, logo, et autres (cf. Figure 22). Les caractéristiques sont évaluées selon quatre directions : horizontale, verticale, diagonale gauche et diagonale droite. Afin de réaliser cette classification, un ensemble de 25 caractéristiques est considéré, parmi lesquelles le *run length*, l'auto-corrélation ou encore le ratio entre la longueur et la largeur des composantes connexes.

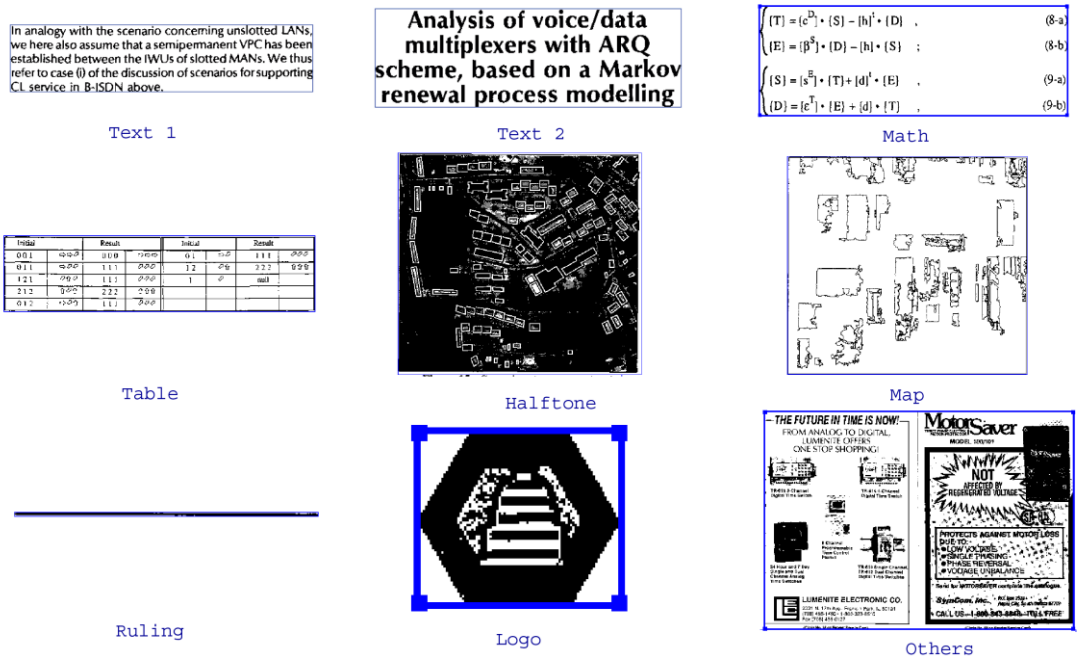


Figure 22 - Illustration de labels pouvant se trouver dans les documents. Illustration extraite de [WaPH06].

Un arbre binaire de classification appliqué sur les caractéristiques étiquette le bloc considéré. Cette approche avait déjà été explorée par la même équipe dans [HaHP95] [WaPH02].

Vauthier et Belaïd [VaBe12] cherchent à analyser la mise en page complète d'un document. Ils s'intéressent à la mise en page logique de l'image comme les adresses, les numéros de téléphone, les formules de politesse, *etc.*, et à la mise en page physique pour identifier cinq types de régions : texte, tableau, signature, tampon, logo. La partie segmentation a déjà été expliquée dans la Section 1.2.2. Concernant la classification, les auteurs utilisent des descripteurs morphologiques (*run length*, composantes connexes, Bi-level Co-occurrence, *etc.*)

sur les zones définies. Ces différentes catégories fournissent 102 caractéristiques dont 43 sont sélectionnés pour ne garder que les plus efficaces grâce à une méthode reprenant l'heuristique de Marc A. Hall [Hall98]. Un classifieur modifié utilisant le *boosting* et les arbres de décision (*PBoost*, une amélioration de *LogitBoost*, dont ils comparent les résultats) est appliqué pour finalement obtenir comme résultat un document étiqueté.

Nous avons déjà évoqué les travaux de Coppi *et al.* [CoGC14] qui utilisent en prétraitement l'algorithme X-Y cut pour segmenter la page d'un document en blanc. Nous allons maintenant nous intéresser à la façon dont Coppi *et al.* classifient ces blocs en utilisant un apprentissage supervisé *via* l'algorithme SVM avec un noyau *Radial Basis Function (RBF)* où chaque bloc est décrit par 308 caractéristiques avec les paramètres fixés par la validation croisée. Leur approche utilise une caractéristique de corrélation locale qui améliore la détection d'un texte dans une image de document.

Keysers *et al.* [KeSB06] cherchent quant à eux à diminuer le nombre d'images nécessaires pour l'apprentissage des zones. Les caractéristiques extraites des blocs sont des histogrammes de texture, de distance, de position invariante, de position relative et des caractéristiques statistiques. Après avoir extrait les caractéristiques, l'approche consiste à les classifier par une méthode des plus proches voisins. La mesure de distance utilisée pour les histogrammes est celle de Jensen-Shannon et pour les autres la distance Euclidienne. Cette étude est particulièrement intéressante car elle offre une comparaison des différents résultats obtenus en évaluant quantitativement des caractéristiques différentes. Elle est également pertinente par le fait qu'elle traite des problèmes liés à une mauvaise acquisition par scanner.

Bouguelia *et al.* [BoBB13] proposent une méthode pour classifier les zones d'une image de document. Ils se placent dans un contexte industriel où ils doivent traiter des documents administratifs. L'étape de segmentation est la même que dans [VaBe12]. Chaque zone obtenue est représentée par un vecteur de caractéristiques comme *run length*, bi-level de co-ocurence et composantes connexes (taille et densité). Chaque zone est classée dans une classe particulière c'est-à-dire logo, tableau, imprimé, *etc.* La classification contient une classe "rejet" pour exclure des zones ambiguës. C'est ensuite à un expert humain de classer la zone. La classification se fait de manière incrémentale. En parallèle, cette méthode peut servir à classifier les documents entre eux. Cette méthode a été testée sur des documents administratifs et d'après les résultats présentés, le taux de reconnaissance atteint 92%.

➢ Conclusion

Les méthodes de classification des régions vues précédemment reposent sur des caractéristiques de textures et des caractéristiques géométriques. La plupart des méthodes se basent sur des caractéristiques communes comme les *run length* ou les caractéristiques extraites des composantes connexes. Ces méthodes se différencient surtout par leur méthode de classification. Elles se fondent généralement sur des caractéristiques statistiques et utilisent un apprentissage supervisé. Ces différences ont été résumées dans le Tableau 3 où nous observons que contrairement aux méthodes de classification fondée pixels le nombre de labels détectés est plus important.

Tableau 3 - Synthèse des principales méthodes de classification des régions dans les images de documents.

Méthodes	Algorithme de segmentation	Type de caractéristiques	Nombre de caractéristiques	Type de classifieurs	Classes / Labels
[WaPH06]	Non renseigné	Texture	25	Arbre binaire	- texte - formules mathématiques - tableaux - halftones - cartes - dessins - traits - logos - autres
[VaBe12]	Segmentation par analyse des composantes connexes	Texture et géométriques	43	Boosting	- tableau - signature - tampon - logo
[CoGC14]	X-Y cut	Texture	308	SVM	- texte - illustration
[KeSB06]	Non renseigné	Texture	comparatif	KNN	- texte - formules mathématiques - tableaux - *halftones* - dessins - traits - logos - autres
[BoBB13]	Segmentation par analyse des composantes connexes	Texture et géométrique	101	KNN	- logos - tableaux - imprimés - manuscrit - signature

Ces méthodes ont l'avantage de donner des résultats très satisfaisants, tout en pouvant extraire de nombreux labels. Mais la dépendance de ces méthodes avec l'étape de segmentation

est un inconvénient, tout comme l'apprentissage qui doit se fonder sur un grand volume de données. Une erreur de segmentation pourra entraîner un problème de classification.

De manière générale les méthodes d'apprentissage demandent d'apprendre sur un grand nombre de données. Ces méthodes sont dépendantes de la nature des documents qu'elles peuvent traiter. Nous verrons par la suite comment certains ont restreint le problème en recherchant à extraire les éléments relevant d'un label en particulier.

1.4 État de l'art sur les méthodes d'extraction par couches

Dans cette partie, nous présentons les approches par couches qui cherchent, à la différence des méthodes présentées jusque-là dans cet état de l'art, à extraire des éléments bien précis dans le contenu du document. Chaque média (tableau, logo, *etc.*) constitue ici une couche de l'image du document, et sera extraite toujours sans segmentation préalable. Cette approche par couches est dite *Multi-Layer Approach* en anglais. Un très grand nombre d'approches par *layers* ont déjà été proposées dans la littérature. Nous ne mentionnerons ici que quelques méthodes représentatives permettant de rechercher du texte imprimé, du texte manuscrit, des tableaux, des séparateurs, des logos. Nous décomposons cette section en sous-sections correspondant aux différents types de couches / *layers* traitées par les approches présentées.

L'analyse par couches peut se faire à différents niveaux :

- la détection : la couche est présente dans le document à analyser ;
- la localisation : la position des éléments de cette couche est précisée dans la page ;
- la reconnaissance : le contenu de la couche est reconstitué.

Nous allons présenter par la suite les couches imprimé, manuscrit, tableau, et logo. Nous terminerons cette section par une discussion sur l'intérêt de l'extraction de ces couches et de leur type.

1.4.1 Extraction de la couche « Texte »

Le texte est la couche la plus représentée dans les documents, elle est également celle qui porte les principales informations. L'expression des idées par la formulation du langage et donc par l'expression est encore le moyen le plus clair pour éviter une mauvaise compréhension. L'extraction du texte permet également d'améliorer la reconnaissance de caractères. L'état de l'art [BSND18] sur la séparation du texte et non-texte dans les images de documents hors ligne

présente les défis scientifiques de cette séparation et classe les méthodes en fonction de ces défis.

- **Le texte « Imprimé »**

Le texte imprimé est majoritairement présent dans les documents contemporains, car ceux-ci sont pratiquement tous nativement numériques, généralement dans des problématiques de partage ou de sauvegarde. De nombreuses méthodes cherchent à séparer la couche texte (imprimé) du reste. Certaines méthodes de segmentation que nous avons présentées précédemment finissent en séparant le texte du reste (*eg.* [WaWC82], [Bloo91]).

La problématique de la détection de texte n'est pas propre au domaine du document et de la vidéo. De nombreux chercheurs se sont intéressés à la détection de textes dans des images naturelles [ZJYW17]. Ce domaine a été grandement sollicité depuis l'arrivée de la réalité augmentée.

- **La séparation « Imprimé / Manuscrit »**

De nombreuses méthodes recherchant l'imprimé ne sont pas de vraies approches par couches car elles n'effectuent pas seulement une classification en textes manuscrits et autres. Elles cherchent à différentier l'imprimé du manuscrit. Cela introduit avant la recherche des autres couches, un pré-traitement qui isole la couche textuelle (manuscrit et imprimé) des autres.

Il existe de nombreuses méthodes pour séparer le manuscrit de l'imprimé. Il s'agit d'un problème encore ouvert. À l'heure de la rédaction de cette thèse, de nombreux travaux se focalisent sur cette problématique. Dans ce paragraphe nous présenterons quelques-unes de ces méthodes. La plupart de ces méthodes comportent quatre étapes : prétraitement, segmentation de la page, extraction des caractéristiques et classification.

Hamrouni *et al.* [HaCV14] se basent sur trois concepts : la définition d'une échelle d'observation, d'un espace de représentation et d'une règle de décision. Les auteurs opèrent une classification pour chaque pixel défini par 4 caractéristiques (2 pour la linéarité et 2 pour la régularité) informant s'il s'agit de l'imprimé, du manuscrit ou du fond.

Grzejszczak *et al.* [GrRB12] proposent une méthode pour séparer l'imprimé du manuscrit qui se décompose en quatre étapes :

- un prétraitement des images pour supprimer le bruit (pouvant par exemple être lié aux bordures) et l'inclinaison du document, où l'on supprime les composantes trop petites ;
- une segmentation du contenu en "pseudo-mots" par double RLSA ;
- une reconnaissance par séparateur à vaste marge (SVM) de la classe d'appartenance ;
- une post-correction utilisant le contexte spatial pour affiner la segmentation.

Cette méthode permet de traiter un grand nombre de documents différents qui vont de la prescription médicale à la facture annotée et la facture pré-remplie (cf. Figure 23).

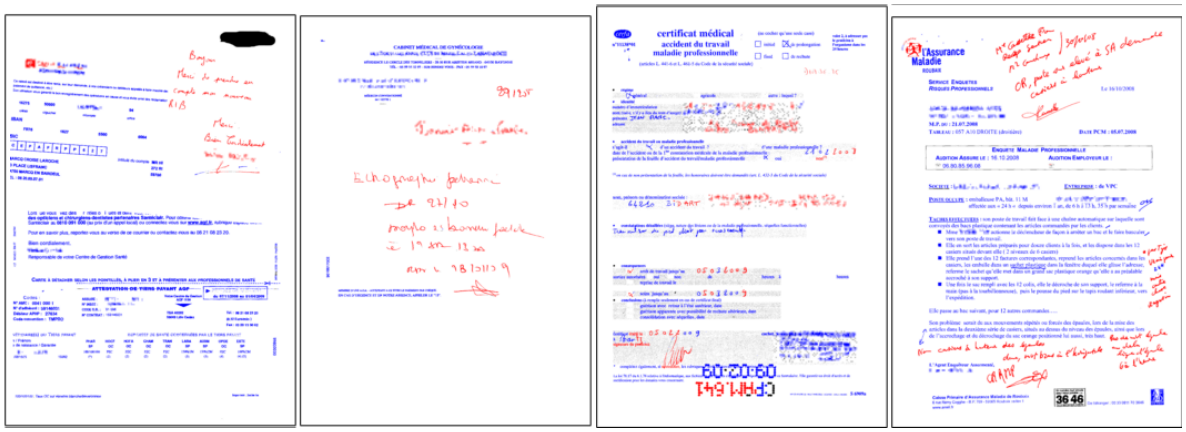


Figure 23 - Illustration des résultats obtenus par la méthode de Grzejszczak et al. (images extraites de [GrRB12]*) (en rouge le texte manuscrit, en bleu : le texte imprimé).*

Ces deux méthodes se distinguent par le nombre de caractéristiques fournies en entrée du classificateur. Au nombre de quatre dans la première méthode, elles sont plus de cent dans la deuxième méthode pour des performances similaires.

1.4.2 Extraction de la couche « Tableau »

Selon le dictionnaire « LAROUSSE »[8] un tableau est, dans le domaine mathématique, un « ensemble d'éléments disposés selon des lignes et des colonnes ou, de façon équivalente, dans les cases d'un rectangle quadrillé. » et dans le domaine de l'imprimerie qui est celui qui nous intéresse une « composition, encadrée ou non, comportant des chiffres et/ou des textes et divisée en colonnes ». Or, face à un monde manipulant une quantité croissante de données, les tableaux sont des outils qui exploitent les capacités visuelles humaines et permettent aux

[8] https://www.larousse.fr/dictionnaires/francais/tableau/76304 visité le 20 mai 2019.

lecteurs de comprendre plus vite et mieux les données présentées. C'est la raison pour laquelle on peut les trouver dans un nombre croissant de documents. Bien restituer un tableau est donc essentiel à la compréhension du document dans lequel il figure. Il est à noter que les tableaux ne sont pas simplement des zones de texte, ils ont une structure particulière qui perturbe l'analyse de la mise en page, que ce soit par leurs séparateurs ou par leurs alignements ou espacements particuliers.

La détection d'un tableau est un problème difficile car les tableaux ont une grande variabilité dans leur mise en page. Les tableaux peuvent avoir ou non des séparateurs matérialisés. S'ils n'en n'ont pas, alors pour faciliter la lecture, les lignes se distinguent souvent par une alternance de couleurs ou par des espacements entre lignes et colonnes plus marqués (cf. Figure 24). De même, l'alignement au sein même du tableau peut être extrêmement variable. Ces caractéristiques variées rendent l'analyse des tableaux très difficile. Les tableaux apparaissant dans un document sont des objets complexes à appréhender et dont les différentes définitions sont ambigües.

	Colonne 1	Colonne 2
Ligne 1	Texte	Lorem Ipsum
Ligne 2	ia ea ob quem	praefus ultimu
Ligne 3	discrimen	Simul sit

(a)

	Colonne 1	**Colonne 2**
Ligne 1	Texte jh	Lorem Ipsum
Ligne 2	ia ea ob quem	praefus ultimu
Ligne 3	discrimen	Simul sit

(b)

	Colonne 1	Colonne 2
Ligne 1	Texte	Lorem Ipsum
Ligne 2	ia ea ob quem	praefus ultimu
Ligne 3	discrimen	Simul sit

(c)

	Colonne 1	Colonne 2
Ligne 1	Texte	Lorem Ipsum
Ligne 2	ia ea ob quem	praefus ultimu
Ligne 3	discrimen	Simul sit

(d)

Figure 24 - Exemples de types de tableaux avec des séparateurs matérialisés ou non. (a) Tableau entièrement matérialisé. (b) Tableau matérialisé par une alternance de couleurs. (c) Tableau semi-matérialisé. (d) Tableau non matérialisé.

On trouve dans la littérature différentes approches de détection et de localisation des tableaux. Les stratégies implémentées par ces approches varient généralement en fonction d'a *priori* sur la façon dont les tableaux sont structurés au sein des images de documents. Les tableaux les plus simples à extraire, et également les tableaux les plus courants, sont ceux qui possèdent des séparateurs matérialisés par des traits. De nombreuses méthodes se fondent sur cette caractéristique.

Cesarini *et al.* [CMSS02] utilisent ainsi la détection de lignes parallèles combinée à une approche par *MXY tree* pour localiser les tableaux. Une méthode proposée dans [RCVF03] permet d'extraire le tableau en récupérant les lignes continues des documents. La méthode introduite dans [GDPP05] repose sur un principe de détection de lignes horizontales et verticales des tableaux ainsi que de détection des points d'intersection.

Cependant, dans de plus en plus de documents, les tableaux ne sont plus forcément matérialisés par des traits (alternance de couleurs, espacements verticaux ou horizontaux, alignements, *etc.*). Les techniques d'extraction doivent alors reposer sur d'autres stratégies et caractéristiques pour détecter et localiser les tableaux. Par exemple, la méthode T-Recs [KiDe01] prend en entrée les mots segmentés puis s'intéresse aux alignements de ces derniers. Mandal *et al.* [MCDC06] utilisent l'espace régulier entre les colonnes.

Dans l'état de l'art sur l'extraction des tableaux [ShSm10], les auteurs ont constaté que la plupart des méthodes de détection de tableaux fonctionnent mal lorsque les documents ont une mise en page de type multi-colonnes

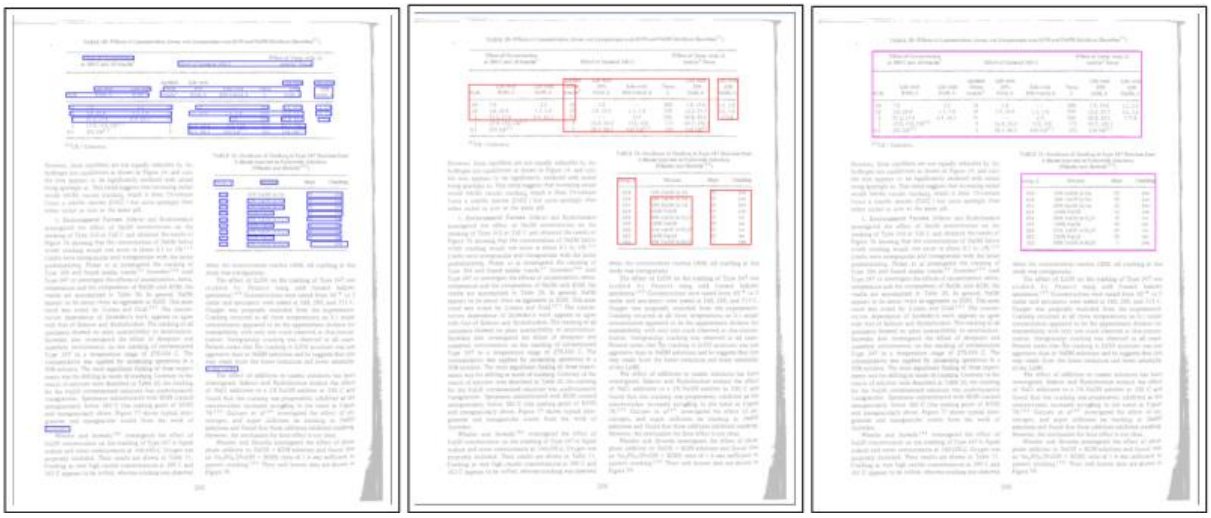
En 2004, Zanibbi *et al.* ont écrit un état de l'art sur l'extraction des tableaux [ZaBC04] dans lequel la structure logique est différenciée de la structure physique des tableaux. La structure physique comprend la localisation des tableaux et des cellules tandis que la logique comprend les types de ces régions et comment ils forment un tableau avec les entêtes de ligne ou de colonne.

Shafait et Smith [ShSm10] cherchent à détecter les tableaux sur une grande variété de documents (rapports d'entreprise, articles de journaux, pages de magazines, *etc.*). L'objectif de leur méthode est de repérer avec précision l'endroit où se trouvent les tableaux et non pas de les analyser. L'avantage de cette méthode est qu'elle fonctionne sur des documents à multiples colonnes. Cette méthode se base sur l'analyse de la mise en page *via* les *tab-stops* qui représente le commencement ou la fin des lignes, méthode qui a été détaillée dans le paragraphe 1.2.3. Leur analyse de la mise en page a révélé que, lorsqu'il y a des tableaux, deux scénarios pour les colonnes de tableaux se dégagent :

- les colonnes du tableau sont considérées comme les colonnes de la page. Cela se produit généralement quand les colonnes sont très bien alignées (l'alignement provoque un grand nombre de tab-stops à détecter. Ils sont assez nombreux pour signaler la présence d'une colonne) ;
- les colonnes de tableaux sont ignorées par le système en raison de cellules qui ne sont pas bien alignées. Par conséquent, la structure en forme de colonnes de la page est correctement identifiée.

Les tab-stop qui pourraient faire partie d'un tableau, c'est-à-dire les lignes particulières qui ont un grand écart entre leurs composantes connexes, qui se composent d'un seul mot (pas d'écart significatif entre les composants), qui se chevauchent le long de l'axe des ordonnées au

sein de la même colonne, sont sélectionnés. Ces lignes permettent, par regroupement, d'obtenir les colonnes et ainsi de récupérer les tableaux (cf. Figure 25).



(a) (b) (c)

Figure 25 - Illustration des résultats obtenus dans [ShSm10]. *(a) Partition des tableaux candidats. (b) Colonnes des tableaux. (c) Régions des tableaux.*

Le *deep learning* a également été utilisé pour détecter les tableaux. En 2016, Hao *et al.* [HGYT16] présentent une méthode de *deep learning* pour détecter les tableaux dans les documents PDF. Des zones pouvant appartenir à un tableau sont sélectionnées sur la base de règles approximatives. Ils utilisent un réseau de convolution se fondant sur des caractéristiques visuelles (comme les caractères et les instructions de rendu) et les caractéristiques non visuelles contenues dans les PDF pour déterminer si les zones sélectionnées sont des tableaux ou non. Puis en 2017, Gilani *et al.* [GQMS17] ont proposé une méthode donnant comme entrée de l'algorithme de *deep learning, Faster Recurrent Convolutional Neural Network (Faster R-CNN),* une image à 3 canaux représentant la transformée en distance Euclidienne, en distance Linéaire et la distance maximum entre les deux [BGKW95], [FCTB08] et [Ragn93], ces transformée étant calculées sur l'image binaire. Cette méthode a montré de meilleurs résultats que la méthode des tab stop [ShSm10].

1.4.3 Extraction de la couche « Séparateur »

Plus que l'extraction de la couche « Séparateur », nous nous intéressons ici à la recherche de droites et de segments, éléments qui constituent généralement les séparateurs. Les droites sont des objets importants dans les documents. Celles-ci permettent de structurer le document comme dans les tableaux ou sont simplement utilisés pour donner une séparation nette entre deux parties. Ils sont généralement constitués d'un ou de plusieurs traits fins mais quelquefois les auteurs des documents se permettent quelque fantaisie en utilisant des images

de traits plus ou moins complexes pouvant être composées de différentes courbes créant ainsi entre autres de jolies arabesques. Dans cette partie nous considérons uniquement les séparateurs constitués de traits rectilignes, c'est pourquoi nous présenterons en premier la méthode de Hough, cette méthode étant l'une des plus connues pour extraire les droites dans une image.

➢ Transformée de Hough

La transformée de Hough (*Hough Transform* (HT)) a été modélisée en 1962 par Hough [Houg62]. Depuis, plus de 2500 articles ont été publiés sur ses variantes, généralisations, propriétés et applications [MuCh15]. Cette transformée permet de détecter les droites dans une image. Cette méthode s'accompagne généralement d'une pré-étape de détection des contours, utilisant généralement le filtre de Canny [Cann86].

Une droite peut être définie de nombreuses manières. Le premier type d'équation que l'on apprend à l'école se réfère aux ordonnées et abscisses à l'origine, on obtient l'équation (1) associée à la Figure 26 (a). Mais l'on peut aussi considérer une représentation polaire illustrée par la Figure 26 (b) et qui conduit à l'équation (2). Dans tous les cas, une droite peut être définie par deux variables, $a \in \mathbb{R}$ et $b \in \mathbb{R}$ pour l'équation (1) et $\rho \in \mathbb{R}$ et $\theta \in [0, \pi[$ pour l'équation (2). Ainsi, si l'on veut représenter dans un nouvel espace chaque droite du plan on obtient un espace en dimension 2. Plusieurs espaces de représentation peuvent être définis dans lesquels une droite est représentée par un point. *A contrario,* un point défini par des coordonnées x et y données est représenté dans l'espace (a, b) par une droite et dans l'espace (ρ, θ) par une sinusoïdale. Ainsi, par cette nouvelle représentation, nous pouvons créer un accumulateur qui compte le nombre de points compris dans une droite. Il suffit ensuite de définir une valeur de seuil pour détecter les droites.

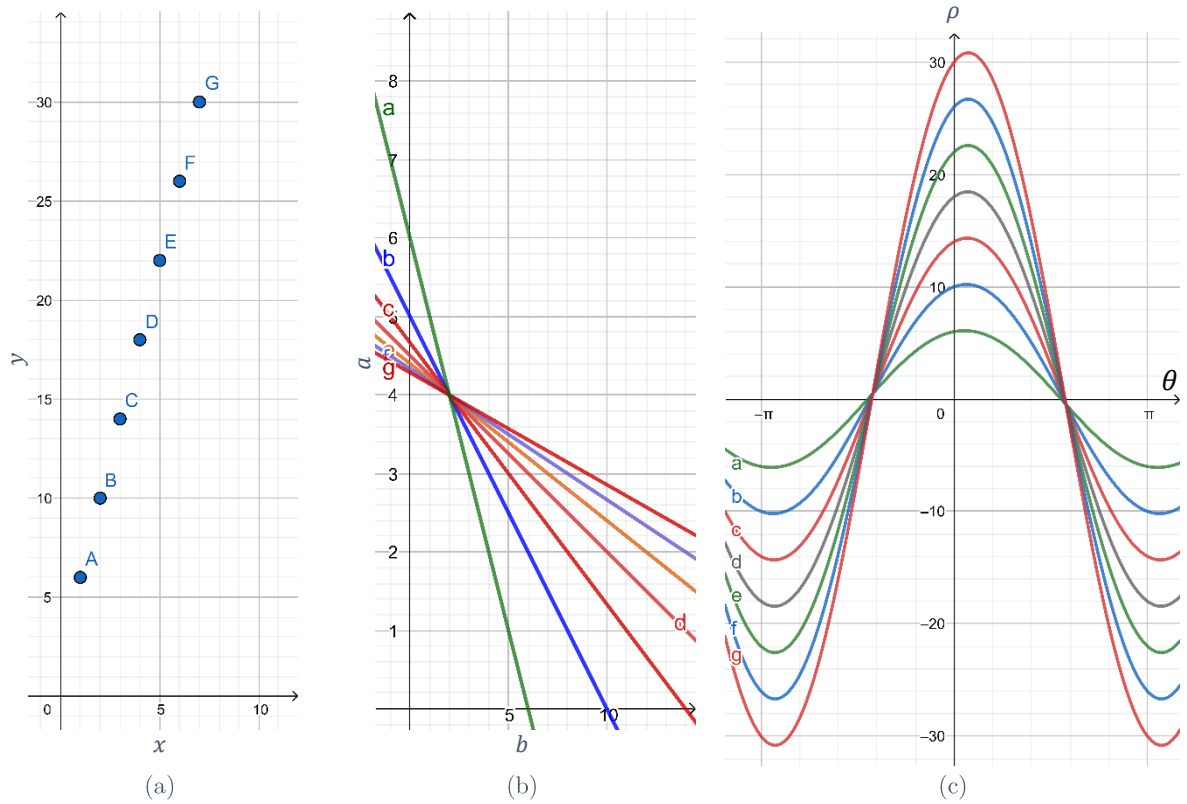
$$y = ax + b \tag{1}$$

$$\rho = x \sin\theta + y \cos\theta \tag{2}$$

(a) (b)

Figure 26 - Caractérisation d'une droite en rouge (a) par ses coordonnées cartésiennes (a, b) et (b) par ses coordonnées polaires (ρ, θ).

Sur la Figure 27 (a), nous pouvons observer 7 points représentés par leurs coordonnées (x, y) et alignés selon une droite de paramètres a = 4 et b = 2 (ou ρ = -0,5 et θ = 1,8). Sur la Figure 27 (b), l'espace de représentation considère que chaque point de l'image (a) comme une droite (de coordonnées $a = \frac{y-b}{x}$), ainsi nous obtenons 7 droites s'intersectant au point de coordonnées (2 ; 4). De même, sur la Figure 27 (c) nous observons 7 sinusoïdales représentant les points de la figure (a) s'intersectant au point de coordonnées (-0,5 ; 1,8).



(a) (b) (c)

Figure 27 - Illustrations de différents espaces de représentation. (a) Image originale. (b) et (c) Espaces de Hough.

De nombreuses variations de cette transformée permettent de réduire le temps de calcul comme *Randomized Hough Transform* (RHT) [XuOK90] qui ajoute de l'aléatoire en ne cherchant plus toutes les droites se trouvant dans une image mais en tirant aléatoirement des couples de points permettant de trouver les paramètres de chaque droite (a et b).

Fast Hough Transform (FHT) [LiLM86] permet également de réduire le temps de calcul en divisant récursivement l'espace de paramètres en hypercubes. Il effectue la transformée de Hough uniquement sur les hypercubes avec des votes dépassant un seuil. La décision permet de savoir si un hypercube reçoit le vote d'un hyperplan si celui-ci intersecte l'hypercube. Cette approche hiérarchique conduit à une réduction significative du calcul et du stockage.

Guil *et al.* [GuVZ95] ont proposé une variante de FHT permettant de détecter les segments en sauvegardant les points qui ont permis de trouver les droites. Ceux-ci sont ordonnés pour trouver les extrémités.

Des variantes de la transformée de Hough permettent de détecter d'autres formes que les droites telles que les familles de courbes, les cercles ou les ellipses (*Generalized Hough Transform* (GHT) [Ball81]).

D'autres méthodes que la méthode de Hough permettent de détecter les droites.

➢ **Autres méthodes**

La méthode de Desolneux *et al.* [DeMM00] compte le nombre de pixels alignés dans une certaine orientation grâce à l'image de gradient et accepte l'ensemble de pixels comme un segment de ligne si la structure observée a une signification apparente. Cette méthode est utilisée comme méthode de validation de ligne. Le défaut de cette méthode est la détection des petits segments, sa nature globale entraine la concaténation de plusieurs petites lignes et augmente le temps d'exécution.

La méthode « EDLines » [AkTo11] permet de détecter les lignes sur les images en niveaux de gris également sans utiliser de paramètre et plus rapide. Cette méthode est constituée de 3 étapes : la première étape détecte les contours grâce à l'algorithme Edge Drawing (ED) [AkTo11]. La deuxième étape cherche dans les contours les alignements des points en utilisant un ajustement de droite par moindres carrés. La dernière étape est une étape de validation des segments. Elle est fondée sur le même principe que celle de [DeMM00]. Cette méthode possède des paramètres fixés fonctionnant pour tout type d'image.

1.4.4 Extraction de la couche « Logo »

Logo est une abréviation de « logotype » qui selon le dictionnaire LAROUSSE, est la « Représentation graphique d'une marque commerciale, du sigle d'un organisme, d'un produit ». Avant d'observer les techniques pour trouver la couche logo, il faut préciser comment est constitué un logo. Un logo peut-être un texte, une image ou un mélange des deux. Un logo sert à identifier un document. Nous pouvons par exemple dire « les documents ayant ce logo concernent cette entreprise ». Un logo devient plus marquant visuellement qu'un simple texte sans typographie particulière. Le proverbe « une image vaut mille mots » illustre bien ce propos.

Certaines approches cherchent à reconnaître les logos déjà enregistrés dans une base de données en cherchant, par exemple à récupérer tous les documents comportant ces logos enregistrés [DoRW96], [SoSa98], [KiKi97], [JaVa98], [MeKW97], [CiSc01], [CFGM97] et [LVTO12]. Cela permet de trier ou de classifier une base d'images de documents où nous pouvons récupérer un document qui nous intéresse.

La méthode présentée par Le *et al.* [LVTO12] permet de reconnaitre des logos dans des documents. Ils commencent par supprimer le bruit par des opérateurs morphologiques. Les points d'intérêt sont extraits par différences de gaussiennes (DoG) puis décrits en utilisant les caractéristiques SIFT. Il y a ensuite une mise en correspondance entre les points trouvés et les points des logos enregistrés dans la base.

Par le même type d'approche les caractéristiques SURF sont en suivant le même principe utilisées par Jain et Doermann en 2012 [JaDo12].

L'inconvénient de ces méthodes est qu'elles sont généralement non supervisées.

1.4.5 Extraction de diverses autres couches

Les documents ne sont pas seulement constitués de textes imprimés, manuscrits, tableaux, logos et de séparateurs. D'autres couches existent, mais il s'agit généralement de documents plus spécifiques. Ainsi, ils peuvent contenir des partitions de musique [CaTV17], des formules chimiques [GhBe16] et [ZaBC02], des équations mathématiques [GYLJ17] ou encore des symboles spécifiques. Les méthodes d'extraction de couches sont liées aux documents que l'on cherche à analyser et aux caractéristiques que possèdent les couches que l'on veut extraire.

Les couches ne sont pas forcément toujours bien définies. Il s'agit souvent, d'identifier le niveau qui nous intéresse. Jusqu'à quelles couches devons-nous approfondir la recherche ? De nombreux articles basent leur recherche sur les couches logiques dans le texte, comme les couches adresses, numéros de téléphone, formules de politesse, *etc.* qui sont des informations logiques [GLEE06].

Les articles cherchant des couches précises sont généralement spécialisés dans un type de documents comme les courriers, factures, chèques de banque, formulaires, *etc.* et peuvent s'appuyer sur les règles de mise en page de ces classes de document. Si nous prenons le cas de la signature, nous sommes en droit de nous demander si celle-ci est une couche à elle seule ou si elle fait partie de la couche manuscrite ou de la couche image. Ces questions sont ouvertes, cela dépend comme toujours du résultat que nous voulons obtenir. Dans cette section, nous

n'avons pas abordé la couche image qui peut être déduite après soustraction des autres couches. L'analyse par couches présente l'avantage d'une extraction indépendante des couches. Nous pouvons donc les utiliser pour rectifier des erreurs, mais en contrepartie, le temps de calcul est plus important.

1.5 Synthèse et discussions

Comme nous l'avons montré dans cet état de l'art, il existe de nombreuses façons d'extraire la mise en page : en segmentant puis en classifiant les zones, en classifiant les pixels ou en extrayant les couches. L'intérêt des méthodes dépend de l'objectif poursuivi et des images sur lesquelles la méthode doit s'appliquer.

Les méthodes de labélisation des régions sont dépendantes de l'étape de segmentation. Dans sa thèse S. Eskenazy [Eske17] a analysé les meilleures méthodes de segmentation selon un critère de stabilité que nous verrons dans le chapitre 4 et a conclu qu'elles n'étaient pas assez stables.

Dans le cadre du projet SHADES, et après réflexions avec les différents partenaires impliqués dans la tâche d'extraction de la mise en page, il nous a fallu faire des choix méthodologiques. Après des expérimentations préliminaires et des comparaisons qualitatives et quantitatives sur les méthodes présentées dans ce chapitre, nous avons constaté que la présence de tableaux dans le document induisait souvent des erreurs dans la phase d'extraction de sa mise en page. En revanche, une fois le tableau extrait, les méthodes sont beaucoup plus efficaces.

Le but du projet SHADES est de travailler sur les documents hybrides. Ce n'est pas un problème courant dans la littérature. Les méthodes présentées ne sont pas axées sur le problème que constitue ces documents. Les méthodes présentées ont été développées dans des cadres différents avec des préoccupations liées aux divers types de documents, alors que nous privilégions la stabilité d'une méthode par rapport aux déformations naturelles qui peuvent se produire au cours du cycle de vie d'un document hybride. Le résultat doit être reproductible si le document s'est dégradé naturellement.

Après analyse des méthodes existantes, nous avons choisi de contribuer à l'analyse de document en créant une nouvelle transformée pouvant être appliquée sur le fond ou la forme pour décrire les documents.

Chapitre 2

L'approche par les lignes : de nouvelles transformées

Sommaire

Résumé

Dans ce chapitre, nous proposons de nouvelles transformées fondées sur les lignes et permettant de décrire une image en fonction d'informations concernant la longueur des lignes et / ou leur orientation. Ces transformées peuvent être définies globalement ou localement en fonction de l'utilisation qui nous intéresse. Nous présenterons également les propriétés que possèdent ces transformées qui ont été définies dans le domaine du continu. De plus, nous présenterons l'efficacité de ces transformées dans le domaine discret, domaine correspondant au domaine de définition des images.

2.1 Introduction

Les primitives traditionnellement utilisées pour décrire et analyser une image de document s'appuient sur trois types d'éléments : les points, les segments (ou les courbes) et les régions [LoBr98]. Il n'y a évidemment pas unicité des décompositions puisqu'une région peut aussi bien être décrite par un ensemble de segments que par un ensemble de points. Notre objectif est ici de choisir un ensemble de primitives qui conduisent à une décomposition unique d'une forme que nous souhaitons représenter puis analyser.

Dans une image, une autre décomposition possible est celle qui oppose le contenu, c'est-à-dire l'objet d'intérêt, du reste de l'image appelé fond. Dans le domaine du document, le contenu principal est le texte, très contrasté par rapport au fond pour en faciliter sa lecture. Nous pouvons noter que des images contenues dans un document sont, sans doute, moins contrastées, dans leur contenu et par rapport au document. Néanmoins, dans ce chapitre nous supposons que l'étape de binarisation qui permet d'extraire ces deux éléments complémentaires a été réalisée et nous considérons des images binaires.

Pour analyser une forme, deux approches sont souvent considérées : calculer des caractéristiques qui seront étudiées d'un point de vue statistique ou utiliser une approche plus structurelle. Cette dernière nous a semblé mieux convenir dans le domaine de l'analyse de documents d'où résultent des images qui possèdent une structure visuelle bien marquée. Dans ce contexte, des primitives ainsi que l'algorithme de décomposition sont donc à définir. Nous avons choisi de limiter les primitives utilisées aux segments de droite, ce qui ne rend pas unique la décomposition d'une image binaire, mais assure une décomposition possible de toute forme tout en diminuant la variété des primitives. Un point (ou un pixel) peut aussi être considéré comme un segment très court (de longueur 1 exprimée en pixels). Les segments sont caractérisés par une origine (un point initial), une longueur et une orientation. À partir d'une décomposition rendue unique par le mode de construction, nous avons caractérisé les formes contenues dans une image binaire.

Dans le cadre de cette thèse, nous avons introduit des transformations d'images permettant d'extraire différentes informations et offrant de nouveaux espaces de représentation pour raisonner. Elles sont basées sur les segments extrémaux contenus dans les formes. Un segment est dit extrémal s'il n'est contenu dans aucun autre segment ayant une longueur supérieure et contenu dans la forme. Deux types d'informations sont considérées pour mieux appréhender, de manière stable, le contenu d'un document :

- d'une part, la longueur des segments ;

- d'autre part, leur orientation.

La longueur des segments de longueur maximale, contenus dans les différents objets (composantes connexes de l'image binaire), portent un certain nombre d'informations pour analyser et caractériser ces formes. Par exemple, une uniformité des longueurs de ces segments indiquera que la forme est plutôt ronde.

Ces transformations d'images constituent une contribution méthodologique de cette thèse qui est relativement générique et applicable à d'autres problématiques de la reconnaissance des formes, potentiellement dans d'autres domaines que l'analyse de documents.

On trouve dans la littérature de nombreuses transformées qui étudient les droites présentes dans une image. On peut citer la transformée de Hough [Houg62] déjà évoquée dans le chapitre 1. Ces transformées modifient l'espace de représentation des formes se trouvant initialement dans le plan cartésien en adaptant le nouvel espace à l'élément « droite » recherché. Cela permet de détecter globalement la présence de droites. Mais dans notre cas, ce sont les segments, plus que les droites que nous recherchons. Pour définir les nouvelles transformées proposées, nous nous référons à la transformée de Radon que nous avons adaptée pour définir une transformée plus locale (cf. Section 2.3).

Les transformées que nous proposons se fondent sur une modification de la perception du document. En effet, nous ne considérons plus l'image comme une matrice où chaque pixel a une valeur représentant son intensité, sa couleur, mais comme un ensemble de droites recouvrant les différentes formes contenues dans l'image. Cela nous permet, entre autres, d'appréhender l'image différemment et dans le cas des images de document, de comprendre ce dernier autrement, par exemple pour y trouver les séparateurs visuels qui structurent les documents. Nous pouvons citer ici, les séparateurs visuels qui séparent le document en colonnes ou encore ceux présents dans les tableaux matérialisés. Par ailleurs, dans cette nouvelle représentation, un caractère sera composé par un ensemble de petits segments orientés dans de multiples directions. De plus, grâce à ces transformées, nous pourrons identifier les éléments en fonction des propriétés des droites mises en évidence.

Ce chapitre s'organise selon le plan suivant : dans la Section 2.2, nous présenterons la transformée de Radon sur laquelle se fonde les transformées que nous proposons, puis dans la Section 2.3 nous verrons comment une modification de la définition peut conduire à l'acquisition d'une information plus locale. S'en suivront la proposition et la description de plusieurs transformées dans la Section 2.4 pour une image binaire et la mise en évidence dans la Section 2.5 de quelques propriétés associées à ces transformées. Ces dernières étant définies dans un

cadre continu, nous aborderons dans la Section 2.6 les implications que le passage au discret induisent, puis nous préciserons l'implémentation que nous en avons faite dans la Section 2.7. Pour finir, nous présenterons une généralisation de ces transformées à des images en niveaux de gris dans la Section 2.8 et nous conclurons ce chapitre dans la Section 2.9.

2.2 Transformée de Radon

La transformée de Radon [Rado17] s'applique sur une fonction $f: \mathbb{R} \times \mathbb{R} \to V$ qui associe à chaque point (x, y) du plan, une valeur $f(x, y)$ appartenant à l'ensemble V. Dans le cadre du traitement d'image, f est associée à l'image et est définie sur un support fini noté Δ_f. Pour une image binaire, l'ensemble d'arrivée utilisé V ne contient que deux éléments 0 et 1 (V = {0,1}). Par la suite, nous convenons que 0 représente le blanc (le fond) et 1 représente le noir (la forme).

La transformée de Radon R d'une image f associe à une droite donnée par ses coordonnées polaires (sa distance à l'origine $\rho \in]-\infty, +\infty[$ et son orientation $\theta \in [0, \pi[$ (cf. Figure 26)) une valeur réelle calculée par l'expression :

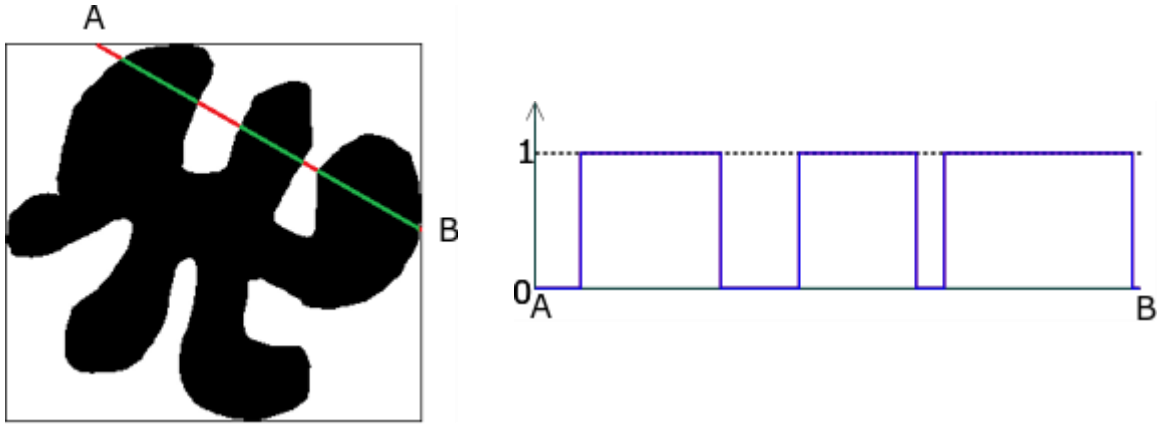
$$\begin{gathered} R: \mathcal{F}(\mathbb{R} \times \mathbb{R} \to V) \to \mathcal{F}(\mathbb{R} \times [0, \pi[\to \mathbb{R}^+) \\ f \to R(f) \\ R(f)(\rho, \theta) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) \times \delta_0\big(\rho - x.\cos(\theta) - \mathrm{y}.sin(\theta)\big)\, \mathrm{dx}\, \mathrm{dy} \end{gathered} \tag{3}$$

avec δ_0 la distribution de Dirac centrée en 0 et $\rho = x.\cos(\theta) + \mathrm{y}.sin(\theta)$ une équation de la droite de caractéristique (ρ, θ). La transformée de Radon d'une image binaire, calculée sur une droite indique la longueur de l'intersection de la forme avec la droite.

La transformée de Radon n'est donc pas définie sur le plan image mais dans l'ensemble des droites du plan. La Figure 28 illustre le processus de calcul, ici f correspond à l'image (cf. Figure 28 (a)), (ρ, θ) désigne une droite représentée par les couleurs verte et rouge (rouge si $f(x, y)$ est égale à 0 et vert sinon). La transformation de Radon de f rend la valeur de l'intégrale de l'image le long de cette droite. La Figure 28 (b) représente l'évolution des niveaux de gris de l'image le long du segment AB. La transformée de Radon calculée pour la droite passant par [AB] est égale à l'aire sous la courbe représentée en bleu sur la Figure 28 (b). Comme l'image est binaire l'aire est aussi la longueur de la partie verte de la droite.

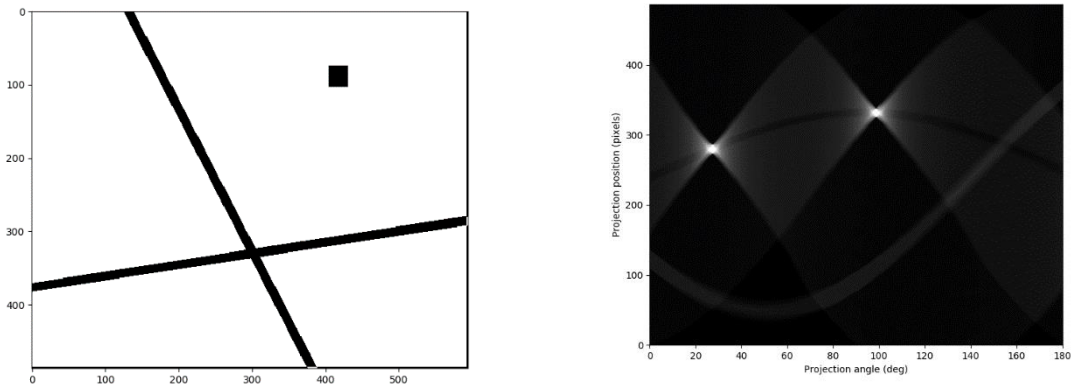
Dans une image binaire, qui par définition est représentée par un signal quantifié, la transformée de Radon équivaut à la transformée de Hough, qui utilise le même changement de

représentation. Chaque point de l'image initiale est représenté dans l'espace de Radon, par une sinusoïde et chaque droite de l'image initiale est représentée par un point dans l'espace de Radon. Sur la Figure 29 (a), une image est représentée contenant deux droites et un pavé de quelques pixels de large. Sur la Figure 29 (b) qui représente le résultat de la transformée de Radon calculé sur l'image (a), on observe des faisceaux de sinusoïdes qui s'intersectent en deux points plus lumineux représentant le nombre de points de l'image qui appartiennent aux deux droites. Le point du milieu, de coordonnées (97, 332) correspond à la droite relativement verticale, l'autre point de coordonnées (35, 380) correspond à la seconde droite et on distingue en bas une sinusoïde assez large correspondant à l'ensemble des points du pavé.



(a) (b)

Figure 28 - Illustration du calcul de la transformée de Radon (globale). (a) Définition d'une droite passant par les points A et B sur une image binaire. (b) Section de l'image suivant le segment [AB].



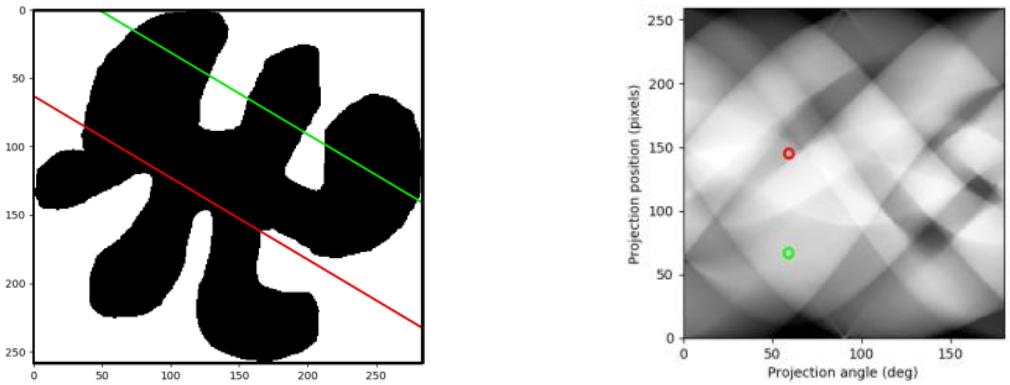
(a) (b)

Figure 29 - Transformée de Radon d'une image comportant deux droites et un pavé. (a) Image initiale. (b) Transformée de Radon de l'image initiale dans l'espace de Radon.

De nombreuses applications ont utilisé la transformée de Radon en traitement d'images et particulièrement en reconnaissance des formes [NaTZ12]. On peut noter des propositions dans le domaine de la reconnaissance des routes [ZhCo07] et des bâtiments [RoEH11] sur des images satellites, des iris [BVMR15], des empreintes digitales [HBSM08], *etc.* Elle a également

été utilisée en traitement d'images médicales, notamment pour la recherche d'images similaires [NTBT16].

La transformée de Radon met en évidence des informations globales sur la présence de droites dans l'image, mais pour l'étude d'un document, nous avons besoin d'une analyse plus locale donnant des informations plus fiables sur une zone locale du document. Sur la forme présente dans la Figure 30 (a), on perçoit six zones reliées par une zone allongée centrale. Il est difficile, sur la transformée de Radon présentée dans la Figure 30 (b), de sélectionner les bonnes droites ou les directions pour caractériser cette forme. Les droites verte et rouge sur la Figure 30 (a) sont représentées par des points de couleur correspondante en (b). Elles n'ont pas le même comportement relativement à la forme, dans un cas (vert) nous avons trois segments, dans l'autre (rouge) il n'y en a qu'un. Par contre dans l'espace de Radon ces droites conduisent à peu près aux mêmes valeurs car la même quantité de forme noire est traversée dans les deux cas. Pour faire face à ce problème, nous proposons, dans le cadre de ces travaux de thèse, une transformée de Radon locale.



(a) (b)

Figure 30 - Transformée de Radon d'une image. (a) Image initiale. (b) Transformée de Radon de l'image initiale dans l'espace de Radon. Les points respectivement rouge et vert dans l'image initiale correspondent aux calculs de l'intégrale sur les droites respectivement rouge et verte.

Cette version plus locale s'appuie sur la définition de la forme telle que nous l'avons définie dans l'introduction (cf. Section 2.1). Celle-ci est ainsi définie sur une image binaire.

2.3 Transformée de Radon locale

Nous proposons une version locale de la transformée de Radon : *Local Radon* (LR). Nous voulons exprimer qu'en un point $P(x, y)$ considérer tous les points de l'image passant par la droite n'est pas forcément intéressant. Seuls les points appartenant à un segment contenant le point P le sont. Dans la Figure 31, nous observons que la droite verte traverse les trois composantes connexes. Pour caractériser le point rouge dans la direction de la droite, il n'est

pas nécessaire de traiter les points appartenant aux composantes connexes se trouvant aux extrémités de l'image.

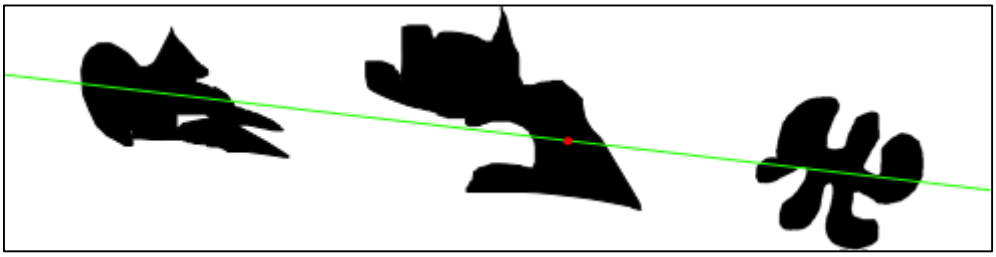


Figure 31 - Illustration d'une droite (en vert) passant par le point rouge et traversant trois formes dans une image.

C'est ce que nous allons exprimer ici. Nous considérons une droite de direction θ et passant par un point de coordonnées $(\rho\ cos(\theta), \rho\ sin(\theta))$. Cette droite est celle qui est définie dans une représentation polaire par le couple (ρ,θ). Sa représentation paramétrique, de paramètre t, est :

$$\begin{cases} x = \rho\cos(\theta) - tsin(\theta) \\ y = \rho sin(\theta) + tcos(\theta) \end{cases} \text{avec } t \in [-\infty, +\infty] \tag{4}$$

La transformée de Radon d'une image f peut alors être calculée comme :

$$\begin{aligned} R(f)(\rho,\theta) = \int_{-\infty}^{+\infty} & f\big(\rho\cos(\theta) - tsin(\theta), \rho\sin(\theta) + tcos(\theta)\big) \\ & \times \delta_0\big(\rho - \rho\cos(\theta) - tsin(\theta)).\cos(\theta) - (\rho sin(\theta) + tcos(\theta)).sin(\theta)\big)\, dt \end{aligned} \tag{5}$$

Partant de cette définition, nous proposons de définir une valeur relative à chaque point P de l'image, nous lui affectons alors une nouvelle caractéristique dépendante du contenu de l'image. Notons (x_0, y_0) les coordonnées de P. Les droites passant par ce point sont caractérisées par la paire (ρ, θ) où ρ et θ sont liés par la relation :

$$\rho(\theta) = x_0\cos(\theta) + y_0 sin(\theta) \tag{6}$$

Ainsi, sur chaque ligne caractérisée par $(\rho(\theta), \theta)$ passant par le point P, le paramètre t caractérisant le point P sur la droite est défini par $t_0 = \frac{x_0 - \rho\cos(\theta)}{-sin(\theta)} = \frac{y_0 - \rho\, sin(\theta)}{\cos(\theta)}$.

En chaque point, la transformée de Radon locale que nous noterons LR est alors définie par :

$$\begin{aligned} LR: \mathcal{F}\big(\Delta_f \to V\big) &\to \mathcal{F}([0,\pi[\times \Delta_f \to \mathbb{R}^+) \\ f &\to LR(f) \end{aligned} \tag{7}$$

$$LR(f)(\theta, x_0, y_0) = \int_{-\infty}^{+\infty} f(\rho(\theta)\cos(\theta) - tsin(\theta), \rho(\theta)\sin(\theta) + tcos(\theta)) \times \delta_0\left(\int_{t_0}^{t} f\big(\rho(\theta)\cos(\theta) - usin(\theta), \rho(\theta)\sin(\theta) + ucos(\theta)\big) - 1\, du\right) dt$$

Dans cette expression, le Dirac nous permet de ne retenir dans l'intégration en t que les points de paramètre t pour lesquels le segment de la droite de direction θ des points de paramètre compris entre t_0 et t est dans la forme, c'est-à-dire des points pour lesquels f prend la valeur 1.

Cette transformée (LR) met donc en évidence un voisinage local qui s'adapte à la forme et qui n'est pas la classique intersection de la forme avec un élément structurant de rayon fixé au préalable. Nous avons illustré sur la Figure 32 cette modification. Ici la transformée prend comme valeur pour chaque direction et en chaque point la valeur de l'intégrale sur une portion connexe de la forme dans laquelle ce point est contenu. Dans cette figure, la transformée de Radon locale au point (en rouge) est égale à l'intégrale le long du segment vert.

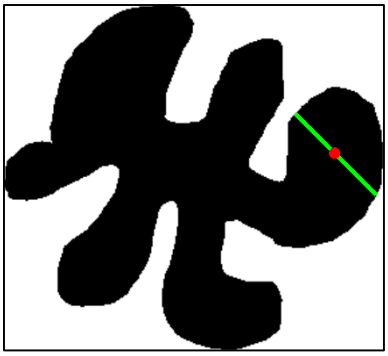


Figure 32 - Illustration du calcul de la transformée de Radon locale (correspondant au calcul de l'intégrale sur le segment vert) définie en un point (en rouge) et dans une direction donnée égale à 45°.

Cette définition nous permet d'introduire les transformées que nous proposons, se fondant sur la définition de ce que nous nommerons un diamètre local de l'objet et conduisant à une nouvelle représentation de la forme intégrant des informations locales.

2.4 Transformées utilisant le diamètre local des objets

Nous avons exprimé dans les paragraphes précédents ce qui nous semble pertinent pour caractériser les formes d'une image. Nous allons dans cette section présenter quatre transformées que nous avons définies, les deux premières permettent d'extraire une information de longueur, tandis que les deux suivantes extraient une information d'orientation. Ces contributions ont fait l'objet d'un article publié à ICDAR en 2017 [ACKO17].

2.4.1 Transformée en diamètre local

Une première transformée a été nommée « transformée en diamètre local (*Local Diameter Transform* (LDT) ». Celle-ci est fondée sur la transformée de Radon locale. Elle définit en chaque point P une information sur la longueur du plus long segment inclus dans la forme qui contient P, par la suite nous nommerons ce segment et sa longueur par le terme de diamètre local au point P. Cette étape est illustrée sur la Figure 33 où les différents segments en vert passant par le point rouge sont matérialisés. Le plus long segment, représenté en jaune, donne la valeur associée à ce point, par la transformée que nous définissons par la suite.

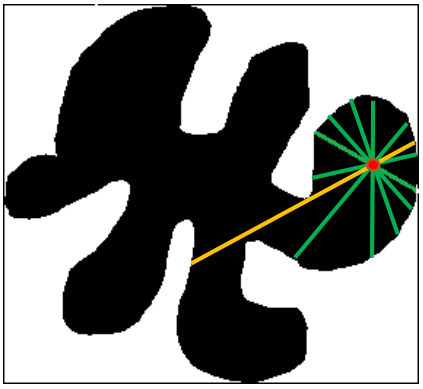


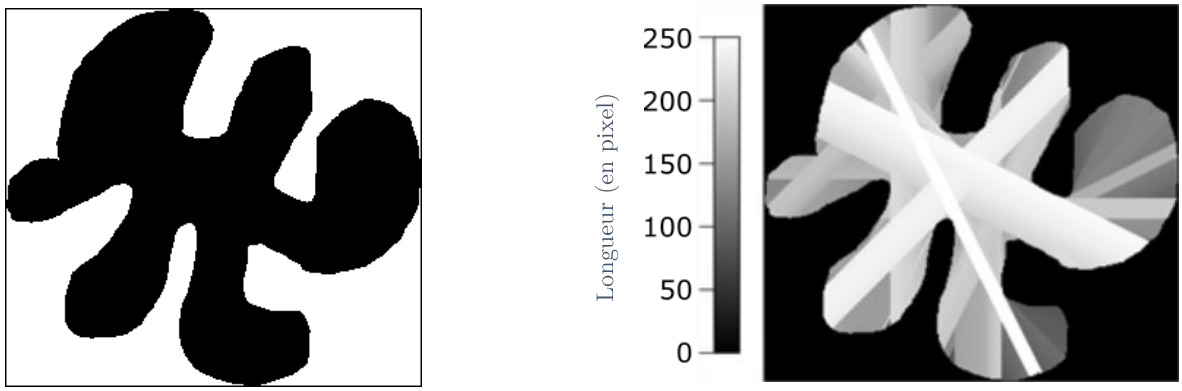
Figure 33 - Illustration du calcul de la Transformée en diamètre local (correspondant au segment jaune) définie en un point (en rouge) et de quelques exemples de segments (en vert) passant par ce point.

Si la transformée de Radon est définie pour chaque droite, nous obtenons ici une transformée définie en chaque point de l'image à analyser. Dans notre espace de représentation, un point est caractérisé par l'ensemble des droites passant par ce point. Le problème est alors de synthétiser toutes les informations fournies par ces directions. Une des informations qui nous a semblé intéressante dans le cas des images, est de considérer la valeur maximale des segments passant par ce point. Il est à noter que la transformée que nous proposons ne fait pas appel à un changement d'espace de représentation comme le faisaient les transformées de Radon et de Radon locale. Le domaine de la transformée est le domaine spatial de l'image initiale, les valeurs associées aux pixels sont des longueurs, donc des réels positifs.

Cette transformée est définie en chaque point de l'image f de coordonnées $(x, y) \in \Delta_f$ par :

$$\begin{aligned} LDT : \mathcal{F}(\Delta_f \to V) &\to \mathcal{F}(\Delta_f \to \mathbb{R}^+) \\ f &\to LDT(f) \\ LDT(f)(x,y) &= \max_{\theta \in [0,\pi[} LR(f)(\theta, x, y) \end{aligned} \tag{8}$$

Le résultat de cette transformée est une image de même dimension que l'image initiale que l'on peut interpréter comme une carte des longueurs des diamètres locaux calculés en chaque point. La Figure 34 illustre le résultat de cette transformée calculée selon 8 directions (cf. Section 2.6). Sur la Figure 34 (a), nous observons que la zone allongée que nous avions mentionnée dans la description de la forme correspond, sur la Figure 34 (b), à une zone où les segments possibles sont très longs, puisqu'ils sont proches de la couleur blanche. Mais la transformée met aussi en évidence des zones claires, que nous pouvons interpréter comme des zones où le diamètre local est important. Nous observons dans la zone en bas à droite la plus grande concentration de pixels sombres. Cela s'explique par son isolement par rapport au reste de la figure. Nous observons également que les plus grands diamètres locaux recouvrent les autres, certains diamètres locaux ne sont maximums que sur une petite partie de leurs pixels, ce qui fait que nous les distinguons peu.



(a) (b)

Figure 34 - Illustration de la transformée en diamètre local (LDT) appliquée sur une image. (a) Image initiale de dimensions 259 × 285 pixels. (b) Résultat de la LDT calculée selon 8 directions : 30°, 45°, 60°, 90°, 120°, 135°, 150°, 180°.

Grâce à cette transformée nous connaissons la longueur du diamètre local associé à chaque pixel, mais nous n'avons pas d'indication effective sur l'importance de ce segment dans la globalité de l'image. Or une telle information est souvent requise dans de nombreux cas pratiques. Une phase de normalisation est alors nécessaire. Le plus souvent, cette dernière est effectuée en comparant les grandeurs : les unes par rapport aux autres, ou à une valeur fixe. On peut considérer par exemple un histogramme de la longueur de ces diamètres sur l'ensemble des pixels en normalisant chaque valeur par le plus grand diamètre. Nous avons préféré une autre normalisation considérant les dimensions de l'image.

Ainsi, s'écartant des normalisations plus classiques, nous présenterons ci-dessous, après avoir précisé la notion d'importance d'un trait, une transformée normalisée prenant en compte l'importance des traits dans une image.

2.4.2 Transformée en diamètre local relatif

Afin d'augmenter l'utilisabilité de cette information et de rendre possible la comparaison de transformées de différentes images. En particulier d'images qui n'ont pas nécessairement les mêmes dimensions. Il nous a paru aussi intéressant de pouvoir comparer les longueurs par exemple des segments de directions différentes. Sur la Figure 35, le segment rouge mesure approximativement la moitié de la hauteur de l'image alors que le segment bleu mesure seulement un sixième de la largeur de l'image. On peut penser que le segment rouge est plus important dans l'interprétation de l'image que le segment bleu. Nous avons voulu rendre cette perception dans une nouvelle transformée déduite de la LDT précédemment définie. Suivant les applications, l'une ou l'autre pourra être choisie.

Les valeurs de la LDT sont normalisées en tout point d'une image en rapportant la valeur de la LDT à la longueur maximum des sections de l'image dans la direction concernée par le diamètre local. Cette valeur sera notée dans la suite par $diam(\Delta_f, \theta)$. Sur la Figure 35 et en appliquant la « transformée en diamètre local relatif » (*Relative Local Diameter Transform* - RLDT), la longueur du segment orange sera normalisée par la longueur du segment vert foncé qui est le plus long contenu dans l'image dans la même direction. On considère ici l'image et non la forme contenue dans l'image.

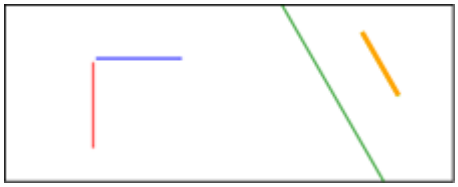


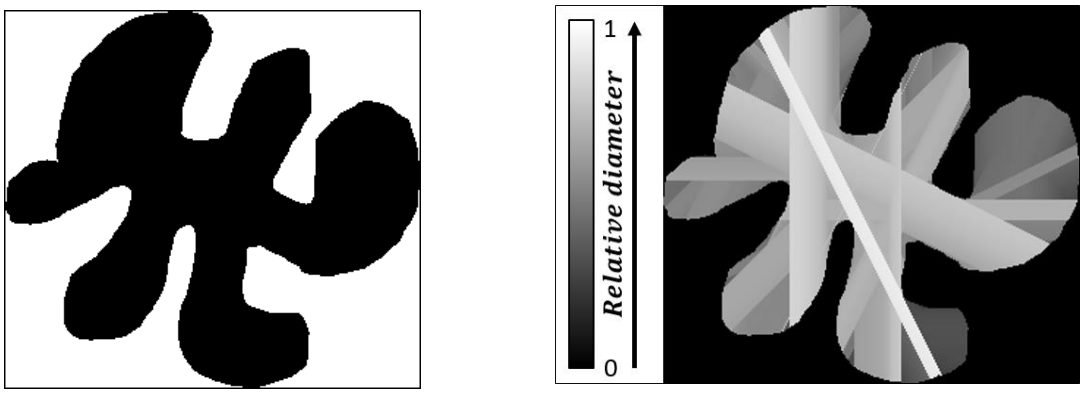
Figure 35 - Illustration de l'interprétation de la grandeur des segments. Le segment rouge mesure approximativement la moitié de la hauteur de l'image alors que le segment bleu mesure seulement un sixième de la largeur de l'image. Le segment orange est grand par rapport au diamètre de l'image dans sa direction représentée par le segment vert.

La transformée en diamètre local relatif est ainsi définie par :

$$\begin{array}{c} RLDT: \mathcal{F}(\Delta_f \to V) \to \mathcal{F}(\Delta_f \to [0,1]) \\ f \qquad\qquad \to \quad RLDT(f) \end{array}$$
$$RLDT(f)(x,y) = \max_{\theta \in [0,\pi[} \frac{LR(f)(\theta,x,y)}{diam(\Delta_f,\theta)} \tag{9}$$

Cette transformée est illustrée par la Figure 36, chaque point étant caractérisé par une distance normalisée selon huit directions (cf. Section 2.6). Pour la visualisation, les valeurs sont normalisées entre 0 et 255. La valeur 255 correspondant à un segment dont les deux extrémités touchent le bord de l'image. Ici les segments à 120° qui traversent la forme dans sa partie

centrale sont moins importants que sur la Figure 36 (b) car ils sont proportionnellement moins grands que ceux verticaux.



(a) (b)

Figure 36 - Illustration de la transformée en diamètre local relatif (RLDT) appliquée à une image. (a) Image binaire initiale de dimensions 259 × 285 pixels. (b) Résultat de la RLDT calculée selon 8 directions : 30°, 45°, 60°, 90°, 120°, 135°, 150°, 180°.

À travers les deux dernières transformées, une information de longueur est obtenue. Mais celle-ci n'est pas la seule information pouvant aider à interpréter les images de documents dans le cadre de notre application. Dans les deux prochaines sections nous présenterons deux transformées permettant d'extraire une autre caractéristique.

2.4.3 Transformée en orientation locale

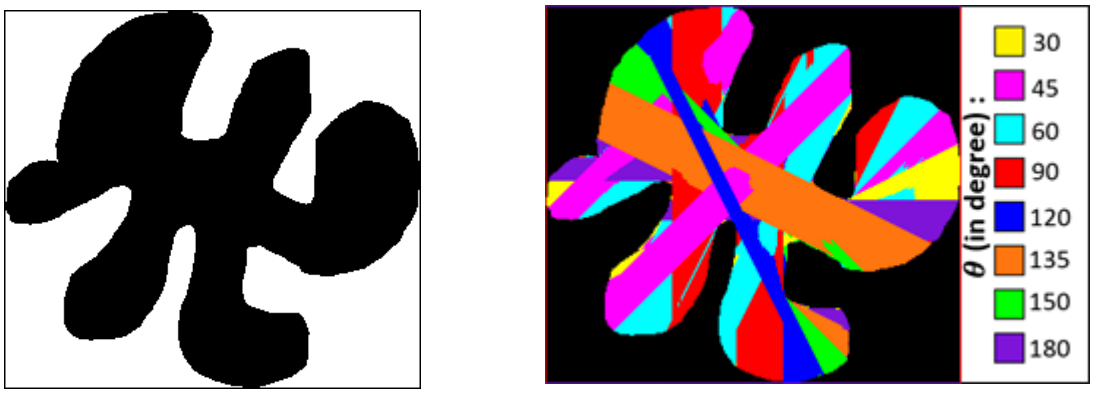
D'autres informations peuvent être extraites des segments que nous venons de définir grâce à la transformée de Radon locale, en particulier l'orientation de ces segments. Cette dernière peut servir à comprendre la nature de la forme étudiée. Par exemple, en traitement d'images de documents, si le document considéré est scanné sans rotation et qu'une composante ne contient que des traits horizontaux ou verticaux, alors il y a une grande probabilité que la composante soit un séparateur horizontal, vertical ou un élément d'un graphique. L'orientation et la longueur sont deux éléments qui contribuent à une bonne perception de la structuration d'un document. Il est alors pertinent de considérer les deux conjointement.

Pour modéliser l'orientation du plus grand segment auquel appartient un pixel, la transformée en orientation locale (*Local Orientation Transform - LOT*) a été définie. Cette transformée, appliquée sur une image, a pour résultat une image de même taille que l'image initiale f où chaque point de Δ_f est caractérisé par la valeur de l'orientation θ ($\theta \in [0, \pi[$), du diamètre local dont la LDT en ce point donne la longueur (cf. Figure 37). Cette transformée est définie par :

$$LOT: \mathcal{F}(\Delta_f \to V) \to \mathcal{F}(\Delta_f \to \mathbb{R}^+)$$

$$f \to LOT(f)$$

$$LOT(f)(x,y) = arg \max_{\theta \in [0,\pi[} LR(f)(\theta,x,y) \quad (10)$$



(a) (b)

Figure 37 - Illustration de la Transformée en Orientation Locale (LOT) appliquée à une image binaire. (a) Image binaire initiale de dimensions 259 × 285 pixels. (b) Résultat de la LOT calculée selon 8 directions : 30°, 45°, 60°, 90°, 120°, 135°, 150°, 180°.

Sur la Figure 37, huit directions sont considérées (cf. Section 2.6). Chaque point est caractérisé par une couleur, chaque couleur étant associée à une des huit directions. Les couleurs sont totalement indépendantes de la longueur des segments considérés. L'aspect des zones de couleur, diffère de celle des zones homogènes dans la représentation de la LDT (cf. Figure 34) ou de la RLDT (cf. Figure 35). Au centre de la figure, plusieurs directions se croisent (les directions 45°, 120° et 135°). Au niveau de la LDT (cf. Figure 34) nous ne pouvions pas voir quel segment était plus grand, ce que l'on peut maintenant distinguer dans la LOT. Sur les pattes de la figure et particulièrement celles de droite, une différence de direction est présente là où les résultats de la LDT étaient globalement homogènes.

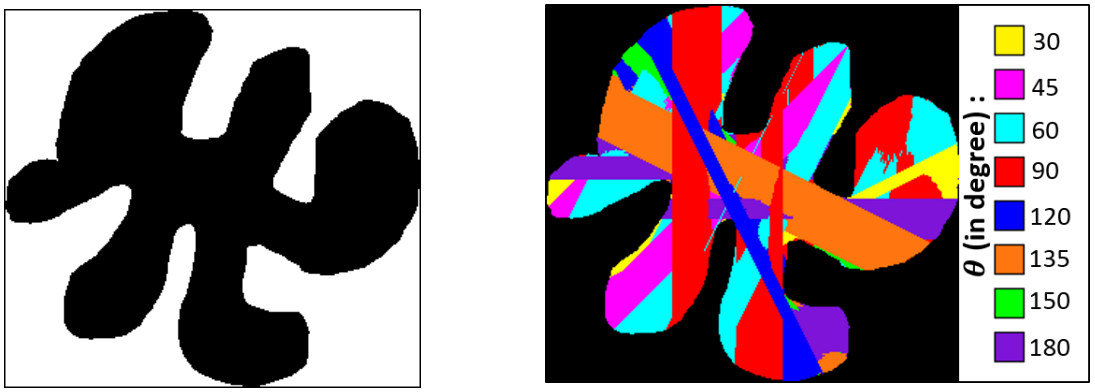
Comme pour la précédente transformation liée à la longueur des diamètres locaux qui a donné lieu d'une part à la LDT et d'autre part à la RLDT, nous avons aussi considéré une version normalisée de cette nouvelle transformée LOT.

2.4.4 Transformée en orientation locale relative

L'information de longueur peut être modélisée en tant que longueur absolue ou en tant que longueur relative. Nous pouvons choisir de garder l'orientation, non pas du diamètre local en un point, mais en considérant l'importance relative des diamètres locaux les uns par rapport aux autres (cf. Figure 38). C'est-à-dire considérer l'orientation du segment ayant la plus grande

importance relativement au domaine Δ_f de l'image f. La transformée en orientation locale relative est donc définie par :

$$RLOT: \mathcal{F}(\Delta_f \to V) \to \mathcal{F}(\Delta_f \to [0,1])$$
$$f \to RLOT(f) \tag{11}$$
$$RLOT(f)(x,y) = arg \max_{\theta \in [0,\pi[} \frac{LR(f)(\theta,x,y)}{diam(\Delta_f,\theta)}$$



(a) (b)

Figure 38 - Illustration de la Transformée en Orientation Locale Relative (RLOT) appliquée à une image. (a) Image initiale de dimensions 259 × 285 pixels. (b) Résultat de la RLOT calculée selon 8 directions : 30°, 45°, 60°, 90°, 120°, 135°, 150°, 180°.

Nous avons présenté de manière théorique, dans cette section, des transformées permettant d'extraire différents types d'informations complémentaires pour des formes définies dans le continu. Ces transformées possèdent plusieurs propriétés que nous allons présenter et discuter ci-dessous.

2.5 Propriétés

Les transformées en diamètre possèdent plusieurs propriétés. Nous allons dans cette section présenter les propriétés des transformées en fonction de modifications spatiales appliquées aux images. Les plus classiques sont la translation, le changement d'échelle et la rotation. L'invariance à ces différentes transformations ou la commutativité de ces transformations avec les transformations que nous venons de définir sont importantes en traitement d'images lors d'opérations géométriques qui peuvent être potentiellement impliquées pendant les acquisitions.

2.5.1 Translation

Soit f' l'image obtenue à partir d'une image f par application d'une translation à la forme suivant un vecteur $\vec{v}$ tel que $f' = tr_{\vec{v}} \circ f$. Soient (x,y), les coordonnées du point P et $(\mathrm{x}',\mathrm{y}')$ les coordonnées du translaté du point P selon le vecteur $\vec{v}$, noté P' :

$$\begin{cases} x' = x + v_x \\ y' = y + v_y \end{cases}$$

Pour les points P contenus dans la forme incluse dans f, la longueur du plus long segment passant par le point P et la direction de celui-ci dans l'image f sont les mêmes que ceux du plus long segment passant par P' dans la forme contenue dans l'image f' :

$LR(f')(\theta, x', y') = LR(f)(\theta, x, y)$ ainsi $LR \circ tr_{\vec{v}}(f) = tr_{\vec{v}} \circ LR(f)$.

Evidemment, on suppose que la translation de la forme, objet de l'étude, et incluse dans f reste contenue dans Δ_f après l'effet de la translation, d'où $\Delta_f = \Delta_{f'}$ et en considérant une direction θ quelconque : $diam(\Delta_{f'}, \theta) = diam(\Delta_f, \theta)$.

Soit pour les transformées :

$$\begin{aligned} LDT(f')(x', y') &= \max_{\theta \in [0,\pi[} LR(f')(\theta, x', y') \\ &= \max_{\theta \in [0,\pi[} LR(f)(\theta, x, y) \\ &= LDT(f)(x, y) \end{aligned}$$

ainsi $LDT \circ tr_{\vec{v}}(f) = tr_{\vec{v}} \circ LDT(f)$. (12)

$$\begin{aligned} RLDT(f')(x', y') &= \max_{\theta \in [0,\pi[} \frac{LR(f')(\theta, x', y')}{diam(\Delta_{f'}, \theta)} \\ &= \max_{\theta \in [0,\pi[} \frac{LR(f)(\theta, x, y)}{diam(\Delta_f, \theta)} \text{ car les diamètres sont égaux} \\ &= RLDT(f)(x, y) \end{aligned}$$

ainsi $RLDT \circ tr_{\vec{v}}(f) = tr_{\vec{v}} \circ RLDT(f)$. (13)

$$\begin{aligned} LOT(f')(x', y') &= arg \max_{\theta \in [0,\pi[} LR(f')(\theta, x', y') \\ &= arg \max_{\theta \in [0,\pi[} LR(f)(\theta, x, y) = LOT(f)(x, y) \end{aligned}$$

ainsi $LOT \circ tr_{\vec{v}}(f) = tr_{\vec{v}} \circ LOT(f)$. (14)

$$RLOT(f')(x',y') = arg \max_{\theta\in[0,\pi[} \frac{LR(f')(\theta,x',y')}{diam(\Delta_{f'},\ \theta)}$$

$$= arg \max_{\theta\in[0,\pi[} \frac{LR(f)(\theta,x,y)}{diam(\Delta_f,\theta)} = RLOT(f)(x,y)$$

ainsi $RLOT \circ tr_{\vec{v}}(f) = tr_{\vec{v}} \circ RLOT(f)$. (15)

Les quatre transformées sont commutatives avec la translation au sens de la composition des transformations. En effet, le calcul de l'une des transformées proposées sur une image suivie de la translation de ce résultat donne la même image de résultats que la translation de l'image initiale suivie du calcul de la même transformée (cf. Figure 39).

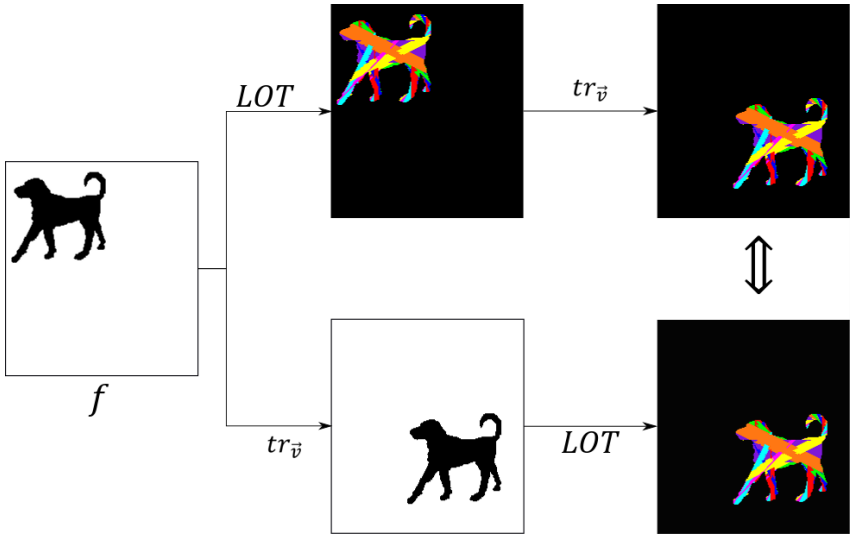


Figure 39 - Illustration de la commutativité de la translation de la LOT.

2.5.2 Changement d'échelle

Lors d'un changement d'échelle (Homothétie), nous pouvons considérer deux types de modifications de l'image :

- la première considère que l'homothétie ($s_{1\lambda}$) ne s'applique que sur la forme et dans ce cas $diam(\Delta_{f'},\theta) = diam(\Delta_f,\theta)$;
- la seconde ($s_{2\lambda}$) considère que la forme dans l'image transformée conserve la même taille que la forme initiale mais globalement le domaine de l'image est modifié.

On suppose dans les deux cas que la modification de la forme ou du domaine de l'image conduit à une forme qui reste entièrement contenue dans le domaine de la nouvelle image.

- **Homothétie sur la forme**

Soit f' l'image de f selon un facteur λ tel que $f' = s_{1\lambda} \circ f$, (x, y), les coordonnées du point P et (x', y') les coordonnées du point P' telles que :

$$\begin{cases} x' = \lambda x \\ y' = \lambda y \end{cases}$$

Ici, on suppose que l'homothétie de la forme est comprise dans l'image et reste contenue dans le domaine de définition de l'image $\Delta_{f,}(diam(\Delta_f, \theta) = diam(\Delta_{f'}, \theta))$. De plus on suppose que $\Delta_f = \Delta_{f'}$.

La longueur du plus long segment passant par le point P' dans l'image f' est le résultat de la longueur du plus long segment passant par P dans l'image f modifiée par le facteur λ tel que :

$$LR(f')(\theta, x', y') = \lambda \times LR(f)(\theta, x, y)$$

Soit pour les transformées :

$$LDT(f')(x', y') = \max_{\theta \in [0,\pi[} LR(f')(\theta, x', y')$$

$$= \lambda \times \max_{\theta \in [0,\pi[} LR(f)(\theta, x, y) = \lambda \times LDT(f)(x, y)$$

ainsi $LDT \circ s_{1\lambda}(f) = \lambda(s_{1\lambda} \circ LDT(f))$. (16)

$$RLDT(f')(x', y') = \max_{\theta \in [0,\pi[} \frac{LR(f')(\theta, x', y')}{diam(\Delta_{f'}, \theta)}$$

$$= \max_{\theta \in [0,\pi[} \frac{\lambda * LR(f)(\theta, x, y)}{diam(\Delta_f, \theta)} = \lambda * RLDT(f)(x, y)$$ car le domaine des images f et f' est resté le même

ainsi $RLDT \circ s_{1\lambda}(f) = \lambda\big(s_{1\lambda} \circ RLDT(f)\big)$. (17)

En conclusion, la LDT et la RLDT ne sont pas commutatives avec l'opérateur $s_{1\lambda}$. Les fonctions $LDT \circ s_{1\lambda}(f)$ et $s_{1\lambda} \circ LDT(f)$ ont le même support (c'est-à-dire sont nulles sur le même ensemble mais les valeurs sont modifiées, et toutes par le même facteur λ. L'homothétie spatiale se traduit par une modification linéaire de la longueur des segments (cf. Figure 40 (a)). Pour les deux autres transformées, nous avons :

$$LOT(f')(x', y') = arg \max_{\theta \in [0,\pi[} LR(f')(\theta, x', y')$$

$$= arg \max_{\theta \in [0,\pi[} \lambda * LR(f)(\theta, x, y) = LOT(f)(x, y)$$

ainsi $LOT \circ s_{1\lambda}(f) = s_{1\lambda} \circ LOT(f)$. (18)

$$RLOT(f')(x', y') = arg \max_{\theta \in [0,\pi[} \frac{LR(f')(\theta, x', y')}{diam(\Delta_{f'}, \theta)}$$

$$= arg \max_{\theta \in [0,\pi[} \frac{\lambda * LR(f)(\theta, x, y)}{diam(\Delta_f, \theta)} = RLOT(f)(x, y)$$

ainsi $RLOT \circ s_{1\lambda}(f) = s_{1\lambda} \circ RLOT(f)$. (19)

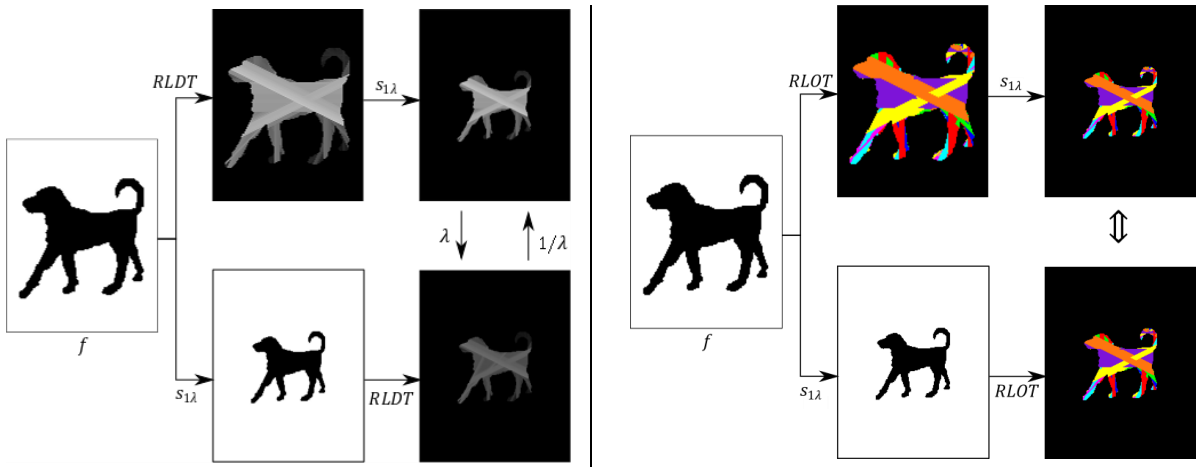


Figure 40 - Illustration du comportement des transformées par rapport à l'homothétie sur la forme. (a) Comportement de la RLDT avec $s_{2\lambda}$. (b) Commutativité de la RLOT avec $s_{2\lambda}$.

La LOT et le RLOT sont commutatives au changement d'échelle car l'ordre entre les longueurs de différents segments est le même que l'ordre entre les homothétiques de ces segments, le plus long segment ne dépend pas de l'échelle considérée. Ainsi, si nous calculons nos transformées en orientation sur une image, puis nous modifions l'échelle sur la forme, le résultat que nous obtenons est le même que si nous changions l'échelle sur l'image initiale puis nous calculions la même transformée (cf. Figure 40 (b)).

➢ Homothétie sur le domaine de définition

Le domaine de définition de f' est ici différent de celui de f, mais ces deux domaines de définitions sont liés par une homothétie. Pour simplifier les notations nous supposerons que le centre de l'homothétie est le centre du domaine Δ_f de l'image f. Ainsi, on a $\Delta_f = \lambda * \Delta_{f'}$. Quelle que soit la direction considérée, nous avons :

$$diam(\Delta_{f'}, \theta) = \lambda * diam(\Delta_f, \theta).$$

Les coordonnées (x, y) d'un point P dans l'image f restent les mêmes dans l'image f' $((x', y') = (x, y))$.

Soit pour les transformées :

$$RLDT(f')(x,y) = \max_{\theta\in[0,\pi[} \frac{LR(f')(\theta,x,y)}{diam(\Delta_{f'},\theta)}$$

$$= \max_{\theta\in[0,\pi[} \frac{LR(f)(\theta,x,y)}{\lambda * diam(\Delta_f,\theta)}$$

$$= \frac{1}{\lambda} RLDT(f)(x,y)$$

ainsi $RLDT \circ s_{2\lambda}(f) = \frac{1}{\lambda}(s_{2\lambda} \circ RLDT(f))$. (20)

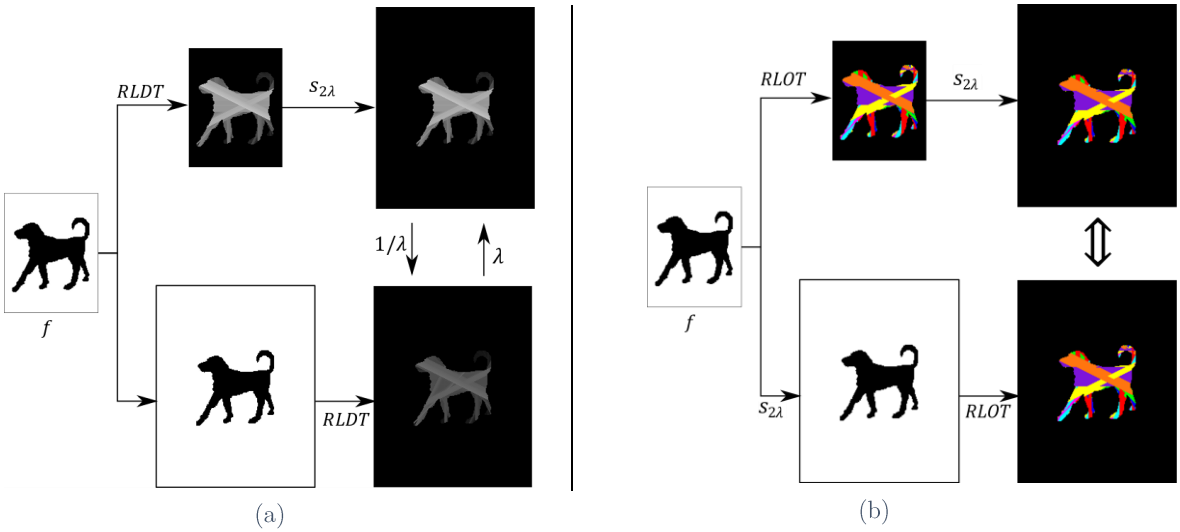
$$RLOT(f')(x',y') = arg \max_{\theta\in[0,\pi[} \frac{LR(f')(\theta,x,y)}{diam(\Delta_{f'},\ \theta)}$$

$$= arg \max_{\theta\in[0,\pi[} \frac{LR(f)(\theta,x,y)}{\lambda * diam(\Delta_f,\theta)}$$

$$= RLOT(f)(x,y)$$

ainsi $RLOT \circ s_{2\lambda}(f) = s_{2\lambda} \circ RL0T(f)$. (21)

Comme pour l'homothétie de la forme, la RLDT est compatible avec un facteur λ avec l'homothétie sur le domaine de définition tandis que la RLOT est commutative avec cette homothétie car la modification homogène du domaine de définition ne modifie pas l'ordre des grandeurs des segments passant par un point (cf. Figure 41).



(a) (b)

Figure 41 - Illustration du comportement des transformées par rapport à l'homothétie sur le domaine de définition. (a) Comportement de la RLDT avec $s_{1\lambda}$. (b) Commutativité de la RLOT avec $s_{1\lambda}$.

La LDT et la LOT ne faisant pas intervenir $diam(\Delta_f, \theta)$sont indépendantes de la taille de l'image et ne dépendent que de la forme elle-même qui n'a pas été modifiée ici.

2.5.3 Rotation

Ici encore, nous supposons que la rotation de la forme, objet de l'étude, et incluse dans f, reste contenue dans Δ_f après l'effet de la rotation d'où $\Delta_f = \Delta_{f'}$.

Soit f' l'image de f après application d'une rotation d'un angle θ' appartenant à $[-\pi, \pi]$ tel que $f' = rot_{\theta'} \circ f$ et (x, y) les coordonnées du point P dans le repère cartésien dont l'origine est le centre de rotation. Les relations reliant les coordonnées cartésiennes du point P avec les coordonnées polaires (r, t) sont :

$$\begin{cases} x = r * \cos t \\ y = r * \sin t \end{cases} \text{où } r = \sqrt{x^2 + y^2} \text{ et t} = 2 \arctan\left(\frac{y}{x + \sqrt{x^2 + y^2}}\right)$$

L'image du point P par la rotation d'angle θ' est le point P' de coordonnées :

$\begin{cases} x' = r * \cos(t + \theta') \\ y' = r * \sin(t + \theta') \end{cases}$ où r et t restent inchangés.

La longueur du plus long segment passant par le point P dans l'image f est la même que celle du plus long segment passant par P' dans l'image f', seul l'angle est modifié :

$$LR(f)(\theta, x, y) = LR(f')(\theta + \theta', x', y') \text{ et } LR(f')(\theta, x', y') = LR(f)(\theta - \theta', x, y).$$

En appliquant le changement de variables (x', y') dans les expressions définissant les transformées, nous pouvons ainsi démontrer l'invariance ou non à un changement de direction. Donc pour la LDT nous obtenons :

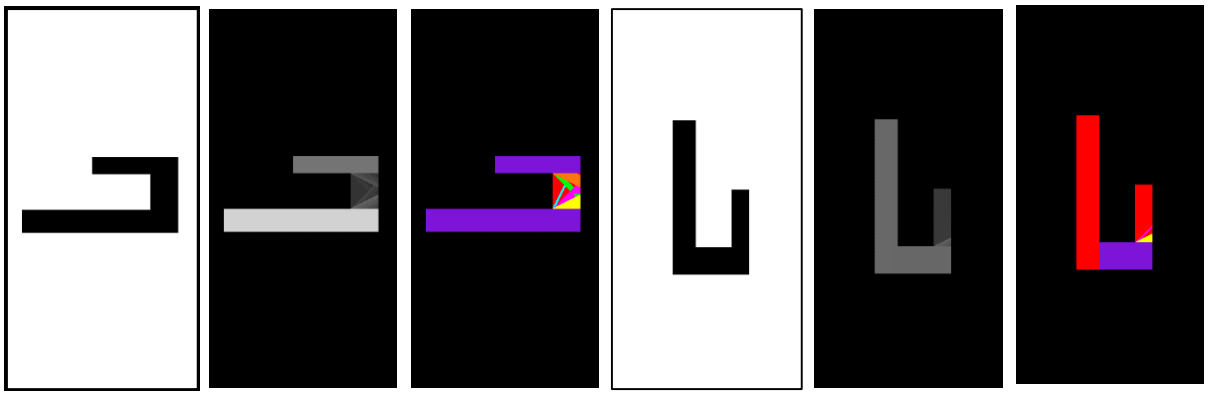
$$LDT(f')(x', y') = \max_{\theta \in [0, \pi[} LR(f')(\theta, x', y')$$

$$= \max_{\theta \in [0, \pi[} LR(f)(\theta - \theta', x, y)$$

soit $\theta'' = \theta - \theta'$ donc $LDT(f')(x', y') = \max_{\theta'' \in [\theta', \pi + \theta']} LR(f)(\theta'', x, y)$ or grâce à la π-périodicité des mesures de direction : $[-\theta', \pi - \theta'[= [0, \pi[$ (π) donc $LDT(f')(x', y') = LDT(f)(x, y)$

ainsi $LDT \circ rot_{\theta'}(f) = rot_{\theta'} \circ LDT(f)$. (22)

Le domaine de définition n'étant pas modifié $(\Delta_f = \Delta_{f'})$, le diamètre de l'image f dans une direction est différent du diamètre dans une autre, $diam(\Delta_{f'}, \theta) \neq diam(\Delta_{f'}, \theta + \theta')$. Les transformées relatives sur l'image f' n'ont donc pas de relation avec celles de l'image f (cf. Figure 42).



(a) (b) (c) (d) (e) (f)

Figure 42 - Illustration des problèmes des transformées relatives avec la rotation. (a) Image initiale. (b) RLDT de l'image initiale. (c) RLOT de l'image initiale. (d) Rotation de l'image initiale selon un angle égal à 90°. (e) RLDT de l'image (d). (f) RLOT de l'image (d).

La transformée en orientation locale n'est pas commutative car le résultat est modifié par la rotation :

$$\begin{aligned} LOT(f')(x',y') &= arg \max_{\theta\in[0,\pi[} LR(f')(\theta,x',y') \\ &= arg \max_{\theta\in[0,\pi[} LR(f)(\theta + \theta',x,y) \\ &= LOT(f)(x,y) + \theta' \end{aligned}$$

ainsi $LOT \circ rot_{\theta'}(f) = t_{\theta'}(rot_{\theta'}(f) \circ LOT(f))$ où $t_{\theta'}$ est la translation du résultat par l'angle θ'. (23)

Nous pouvons conclure que la LDT est commutative avec la rotation, ce qui est logique car le résultat est la plus grande longueur résultat de la *Local Radon* obtenue en considérant toutes les directions. Dans la pratique la propriété n'est respectée que si l'angle de rotation est compatible avec la quantification des angles choisis (cf. Section 2.6).

Dans cette section, nous avons défini et étudié les transformées de fonctions définies sur $\mathbb{R}^2$ et à valeur binaire. Ces fonctions sont à ensemble de définition Δ_f borné. Néanmoins, dans le cas du traitement d'images, l'ensemble de définition de l'image, son domaine, est bien borné mais n'est pas défini sur $\mathbb{R}^2$ mais sur $\mathbb{N}^2$. Dans la section suivante nous allons analyser les contraintes qui en découlent et préciser les choix que nous avons faits pour l'implémentation.

2.6 Du continu au discret

Les images binaires sont une représentation de la réalité dans laquelle le domaine spatial a été discrétisé. Dans la définition des transformées proposées nous avons mis en évidence les

notions de segments et d'orientation. La représentation des droites discrètes est un problème complexe qui a donné lieu à de très nombreuses études [Debl95]. Si dans le domaine continu, nous pouvons appréhender toutes les directions, dans le domaine discret ce n'est pas possible. Considérer seulement les directions également réparties selon les 360 mesures d'angle en degrés est possible dans la réalité (le continu), mais est inadapté à l'étude des contenus d'une image.

Dans le domaine continu, le nombre de segments passant par le centre d'un pixel P est infini. Les implémentations de la transformée de Radon dans le plan discret, recherchent les droites dans une image. Elles considèrent généralement qu'il y a autant de droites passant par un point P de l'image qu'il y a de pixels dans le bord de l'image. La droite discrète est par exemple construite en appliquant un algorithme de tracé de segment de Bresenham [Bres65].

Comme nous nous plaçons d'un point de vue plus local, nous nous intéressons à tous les segments passant par un point P. Dans le domaine continu, le voisinage de taille d d'un point correspond à un cercle de diamètre d, alors que dans le discret, il correspond usuellement à un carré de côte d. En considérant un cercle de diamètre 3 approximé par un voisinage 3×3, nous obtenons quatre directions de droite associées à des segments de longueur 3 (cf. Figure 43). Sur la Figure 43 (c), nous observons ainsi quatre directions possibles 0° (θ_1), 45° (θ_2), 90° (θ_3) et 135° (θ_4). Nous sommes ainsi loin de l'infinité de directions possibles dans le domaine du continu.

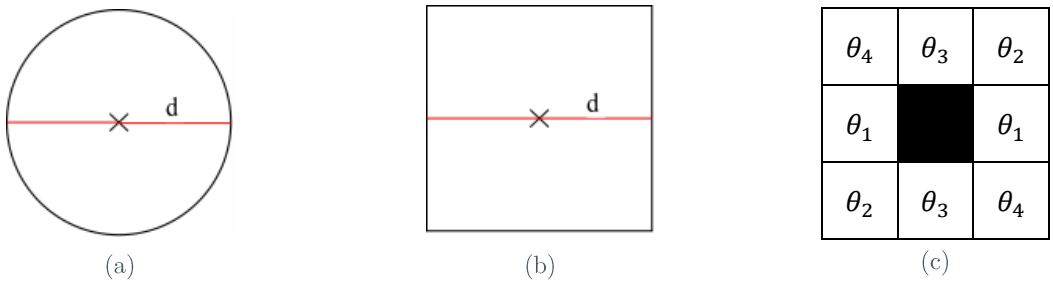


Figure 43 - Voisinage d'un point et discrétisation. (a) Dans le continu selon un voisinage d. (b) Dans le discret selon un voisinage d. (c) Dans le discret selon un voisinage 3×3.

Nous pouvons mettre en évidence deux types d'ambiguïtés concernant la discrétisation des droites, l'une au niveau des pixels des segments et l'autre au niveau de la définition des directions. Selon les applications, ne considérer que quatre directions limite l'information extraite grâce aux transformées que nous avons définies. Nous pouvons également augmenter le rayon du voisinage considéré autour du pixel P. En passant d'un voisinage 3×3 à un voisinage 5×5 de rayon 2, nous augmentons le nombre de directions possibles. La Figure 44, présente 8 directions de droite, θ_0, …, θ_7. Cependant, nous pouvons constater que les droites rouges théoriques traversent les pixels sans forcément passer par leurs centres. De plus, un pixel peut

être traversé par trois segments de directions distinctes. Il s'agit des pixels représentés en jaune. Nous qualifierons ces pixels comme ambigus.

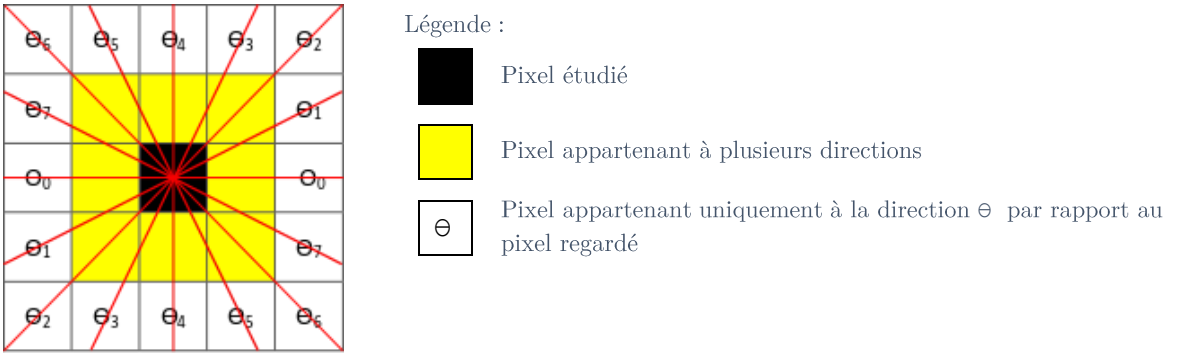


Figure 44 - Illustration des droites discrètes en rouge passant par le centre d'un pixel dans un voisinage 5×5.

La Figure 45 représente huit segments passant par le pixel P. Si le segment est représenté en 8-connexité (et non en 4-connexité) un seul des pixels représentés en orange est suffisant pour définir le segment et il peut être indifféremment l'un ou l'autre. Ainsi, selon la discrétisation du segment, un de ces pixels peut ou non être représenté.

Pour l'analyse d'une image, nous avons choisi de ne considérer que les directions appréhendables dans le voisinage 5×5 d'un point, soit les 8 directions θ_0, …, θ_7 approximant les angles de mesures respectives 0, 30, 45, 60, 90, 120, 135 et 150 degrés. Ce choix est un compromis pour la définition des segments, en termes de temps de calcul et d'erreur d'approximation. Nous pouvons par ailleurs faire deux remarques qui montrent comment des directions intermédiaires peuvent néanmoins être approximées grâce à notre approche sur des voisinages 5×5. La première concerne des segments longs n'ayant pas exactement une des directions modélisées, la seconde remarque concerne l'ajustement de la précision de l'approximation des directions en fonction de l'application.

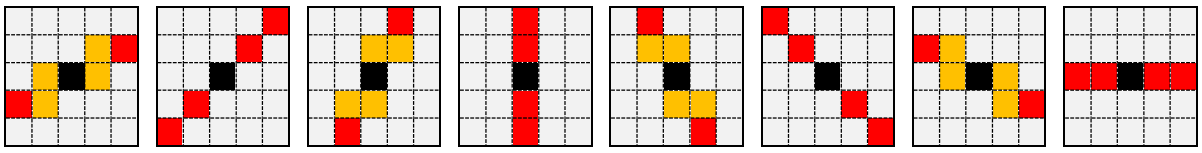


Figure 45 - Caractérisation des segments passant par le centre dans un voisinage 5×5.

L'approximation des 8 directions de manière stricte et locale pourrait laisser penser que de longs segments n'ayant pas exactement les directions modélisées, ne seraient pas extraits. Mais de tels longs segments possédant une direction proche de celles mentionnées précédemment, peuvent être déduits en utilisant la LOT et la LDT en 8 directions comme nous allons le montrer par un exemple. Ainsi, pour des raisons d'efficacité nous avons choisi de ne considérer qu'un nombre réduit de directions. La Figure 46 (a) montre un exemple de segment

(noir) dont la direction est différente mais proche de 0 degré. La méthode considérant les 8 modèles de segments locaux, permet d'affecter à chaque pixel la longueur du plus long segment le contenant. Individuellement les directions associées à chaque pixel sont calculées. Elles sont indiquées dans la Figure 46 (c). L'ensemble de ces segments sont caractérisés par des composantes connexes dans la LOT, de la direction mentionnée (cf. Figure 46 (c)). Ces composantes connexes ont un diamètre (une longueur) différent de la taille indiquée dans la LDT (cf. Figure 46 (b)). Cette manière d'approximer les autres directions peut également être utilisée avec la RLDT et la RLOT.

Nous pouvons également choisir d'analyser une image selon une direction particulière en privilégiant cette direction. Nous pouvons obtenir plus de précision sur les segments qui sont approximativement dans cette direction en modifiant le voisinage considéré. La Figure 47 montre un exemple où la direction horizontale est privilégiée. Le voisinage est étiré dans cette direction dans un rapport qui correspond à la précision voulue. Les points frontières du voisinage définissent les directions, avec un voisinage (3 × (2k+1)), nous définissons 2k directions qui approximent la direction horizontale.

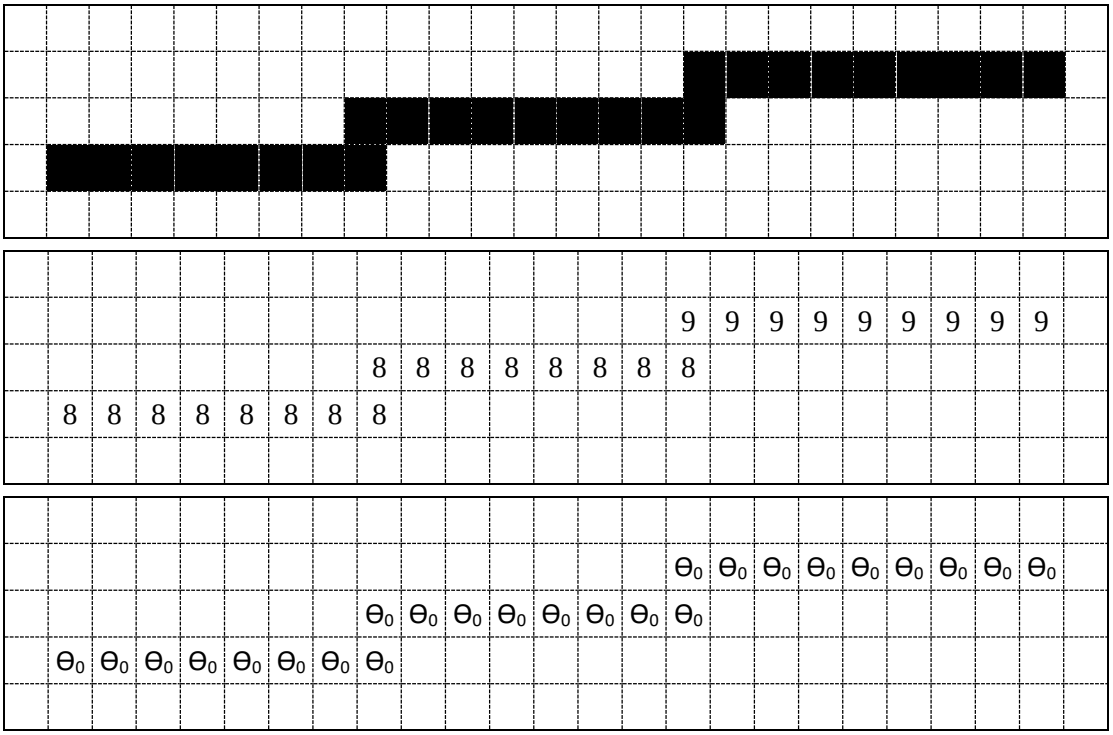


Figure 46 - Segment dont la direction approxime les 180 degrés. (a) Image binaire. (b) LDT. (c) LOT.

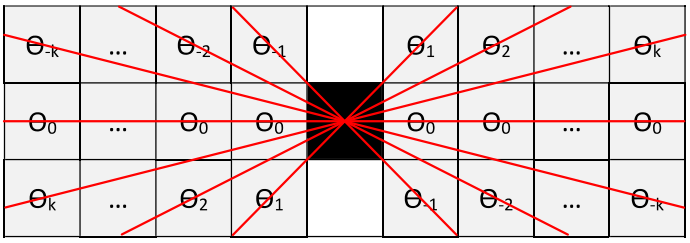


Figure 47 - Approximation de la direction horizontale selon un voisinage (3 × (2k+1)).

Au-delà de ces problèmes liés au passage d'une géométrie dans le continu à une géométrie discrète, nous allons maintenant présenter comment toutes ces transformées ont été implémentées dans notre application.

2.7 Implémentation

Les transformées en diamètre s'appuyant toutes sur la transformée de Radon locale, nous proposons d'en préciser le pseudo-code (**Algorithme 1**) de manière à rendre nos résultats reproductibles.

Tout d'abord nous donnons quelques définitions de fonctions utiles. Nous noterons N le nombre de pixels de l'image binaire I, la fonction « ***Asuivant (p, θ, p')*** » a pour argument un pixel **p** et une direction θ et renvoie le pixel **p'**, le pixel suivant de **p** dans la direction θ en fonction des modèles qui définissent les directions (cf. Figure 45 par exemple). Le résultat de la fonction est un booléen, vrai si le suivant **p'** se trouve dans la forme et faux s'il se trouve dans le fond. En considérant le voisinage défini précédemment (cf. Section 2.6), le pixel suivant d'un pixel **p** est défini comme le pixel à la frontière du voisinage dans la direction donnée θ et ***dist (θ)*** indique la distance euclidienne entre ces deux pixels. Comme cette distance est définie par la direction θ et que le pas entre deux points ne va pas changer, nous pouvons le calculer au début de la fonction. La fonction « ***Ligne*** » prend en paramètre une image, deux points et une valeur entière et trace dans cette image le segment joignant les deux points et leur affecte la valeur donnée.

Algorithme 1 : Transformée de Radon Locale dans la direction θ.

```
Entrées
Image I
Direction θ
Sortie
        Image I_R

Initialisation
        Initialiser (I_R) // initialiser l'image par des 0
        d <- dist(θ)

Chaîne de traitement
        POUR i de 0 à N-1
```

```
                SI I_R[p_i] = 0 et I[p_i] = 1
                        c <- 0
                        p <- p_i
                        TANT_QUE Asuivant(p, θ, p')
                                c <- c + d
                                p <- p'
                        FIN du Tant que
                        Ligne (I_R, p, p_i, c)
                FIN du Si
        FIN du Pour
```

Le diamètre de l'image dans une direction est défini par :

$$diam(\Delta_f, \theta) = \min\left(\frac{nombre\ de\ colonnes\ de\ I}{\cos(\theta)}, \frac{nombre\ de\ lignes\ de\ I}{\sin(\theta)}\right)$$

Les transformées en diamètre peuvent alors être calculées en comparant, pour chaque pixel, les valeurs obtenues dans les transformées de Radon Locale calculées en ce point dans les différentes directions prédéfinies. La LDT garde la plus grande valeur et la RLDT la plus grande valeur normalisée par le diamètre de l'image dans la direction du plus long segment. Tandis que, la LOT conserve la direction dont elle est issue et la RLOT conserve la direction de la plus grande valeur des transformées de Radon locales calculées en ce point, normalisées par le diamètre de l'image dans la direction considérée dans la LR.

En terme de complexité, notre algorithme est très efficace puisque dans notre implémentation, la transformée de Radon locale en une direction θ est 2n (avec n le nombre de pixels de forme de l'image car nous parcourons 2 fois chaque pixel). Par conséquent, chacune des quatre transformées à une complexité égale à cette complexité multipliée par le nombre de directions.

2.8 Du binaire aux niveaux de gris

Jusqu'à présent nous avons considéré dans les parties théoriques précédentes des fonctions à deux valeurs et dans la partie applicative des images binaires dont nous nous sommes donné pour objectif d'étudier les caractéristiques de la forme contenue dans l'image. Dans de nombreuses applications, la forme est issue d'une étape de binarisation d'une image en niveaux de gris. Cette étape conduit bien souvent à une perte d'information qui peut conduire, à cause de la présence de bruit dans l'image, à transformer un trait visible sur l'image en niveaux de gris en une suite de deux segments ou plus. D'autres types de bruit peuvent conduire plusieurs segments à n'en constituer qu'un seul. Pour tendre vers une contribution plus générique, il nous a donc semblé intéressant de généraliser les transformées précédentes même si dorénavant nous ne pourrons plus parler de fond ni de forme dans l'image. Tous les points auront alors le même statut.

Pour pallier ce problème, nous avons étendu la définition de la RLDT aux fonctions à valeurs dans $\mathbb{R}^+$ en ne calculant plus l'intégrale sur une partie de l'image *a priori*, la forme, mais sur l'ensemble des points de l'image. Les transformées en diamètre étant fondées sur la notion de localité de la forme, nous avons cherché à donner un sens à la frontière d'une forme à laquelle chaque pixel pourrait appartenir. Pour cela nous avons ajouté un paramètre β qui permet de gérer l'homogénéité de la couleur des formes. La longueur d'un diamètre local en un point P est calculée en fonction de la différence de niveaux de gris entre le niveau de gris au point P et les niveaux de gris des pixels se trouvant sur la droite de direction donnée. Par exemple, si l'on considère un pixel ayant un niveau de gris égal à 50 et $\beta = 10$, les pixels considérés comme faisant partie du segment passant par P seront les pixels de la droite dans la direction considérée dont le niveau de gris sera compris entre 40 et 60 et qui constitueront une composante connexe contenant P. Les transformées de diamètre en niveaux de gris sont obtenues en remplaçant la fonction LR par $\mathrm{LR}_{\mathrm{nvg}}$ définie par :

$$LR_{nvg}: \mathcal{F}(\Delta_f \to V) \to \mathcal{F}([0,\pi[\times \Delta_f \times \mathbb{R}^+ \to \mathbb{R}^+)$$

$$f \to LRnvg(f)$$

$$\begin{aligned} LRnvg(f)(\theta, x_0, y_0, \beta) &= \int_{-\infty}^{+\infty} f\big(\rho(\theta)\cos(\theta) - tsin(\theta), \rho(\theta)\sin(\theta) + tcos(\theta)\big) \\ &\times \delta_0\left(\int_{t_0}^{t} g\big(u, \theta, x_0, y_0, \beta\big)\ du\right) dt \end{aligned} \tag{24}$$

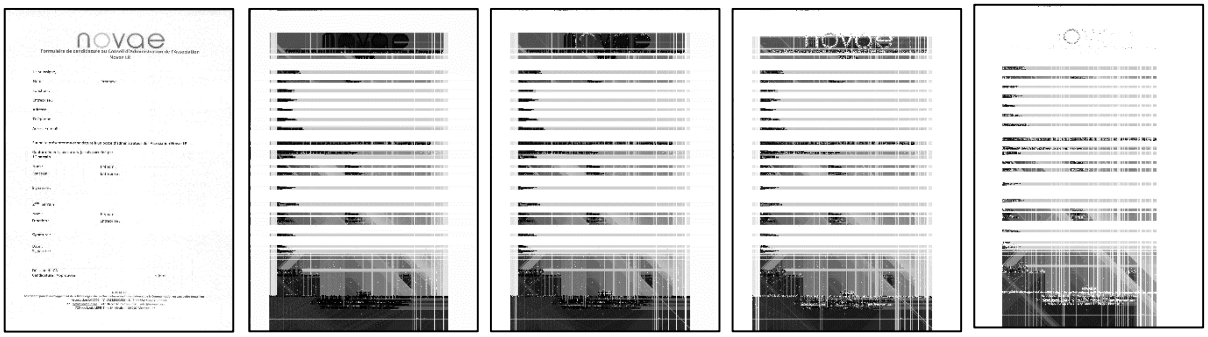
où $g\big(u, \theta, x_0, y_0, \beta\big) = \begin{cases} 0\ si\ |f(\rho(\theta)\cos(\theta) - u.\sin(\theta), \rho(\theta).\sin(\theta) + u.\cos(\theta)) - f(x_0, y_0)| \le \beta \\ 1\ sinon \end{cases}$

Nous observons sur la Figure 48 quelques exemples d'utilisation d'images de transformées en niveaux de gris avec différentes valeurs du paramètre β.La première ligne présente le résultat sur une image de document « propre ». Le seuil β influence peu le résultat, s'il n'est pas trop élevé, peu de différences sont constatées. Le problème vient ici de la difficulté à distinguer le fond de la forme, car même si l'on peut considérer les segments très longs (ceux dont la couleur est claire, proche du blanc) comme faisant partie du fond, nous ne pouvons pas connaître le label des segments courts et une autre information doit être prise en compte. La deuxième ligne contient un document beaucoup plus bruité, le résultat varie en fonction du seuil choisi et, comme dans l'autre image de document, la différentiation du fond et de la forme est un problème. La troisième ligne présente le problème d'une image naturelle, ici un photographe. Nous observons sur les images que le photographe se détache du fond (notamment son manteau) tout en distinguant certains éléments. Il semblerait que considérer plus de directions serait intéressant pour mieux caractériser les objets. La dernière ligne présente une

image satellite. Nous observons sur les résultats de la RLDT en niveaux de gris que les éléments linéaires comme les routes et certaines zones agricoles se détachent du reste de l'image.

L'une des difficultés de l'utilisation de ces nouvelles transformées est que nous ne pouvons plus distinguer le fond de la forme. Ce problème est illustré par la Figure 49 sur laquelle nous avons appliqué la RLDT en binaire (cf. Figure 49 (b)) et en niveaux de gris (cf. Figure 49 (c)). Dans la Figure 49 (a), nous voyons un damier de taille 5×5, avec 13 carrés noirs et 12 carrés blancs :

- dans le premier cas (cf. Figure 49 (b)), nous avons appliqué la RLDT sur l'image binaire en considérant la forme constituée des pixels noirs. Comme sur un carré toutes les longueurs sont relativement identiques, nous distinguons, des lignes plus claires représentant les segments présents dans plus d'un carré. Les pixels noirs dans cette Figure 49 (b) indiquent que l'on est hors de la forme étudiée. Nous pouvons aussi observer que les deux diagonales blanches correspondent à des segments longs traversant l'image ;
- dans le second cas, nous appliquons la RLDT en niveaux de gris à l'image (a). L'image étant binaire, le paramètre β n'a pas d'importance (tant qu'il est inférieur à 255). Les carrés étant de mêmes tailles, nous ne pouvons plus distinguer sur l'image transformée, les contours d'une quelconque forme. Il est alors difficile de dresser une conclusion sur le contenu de l'image, le fond et la forme sont équivalents dans cette image.



(a) (b) (c) (d) (e)

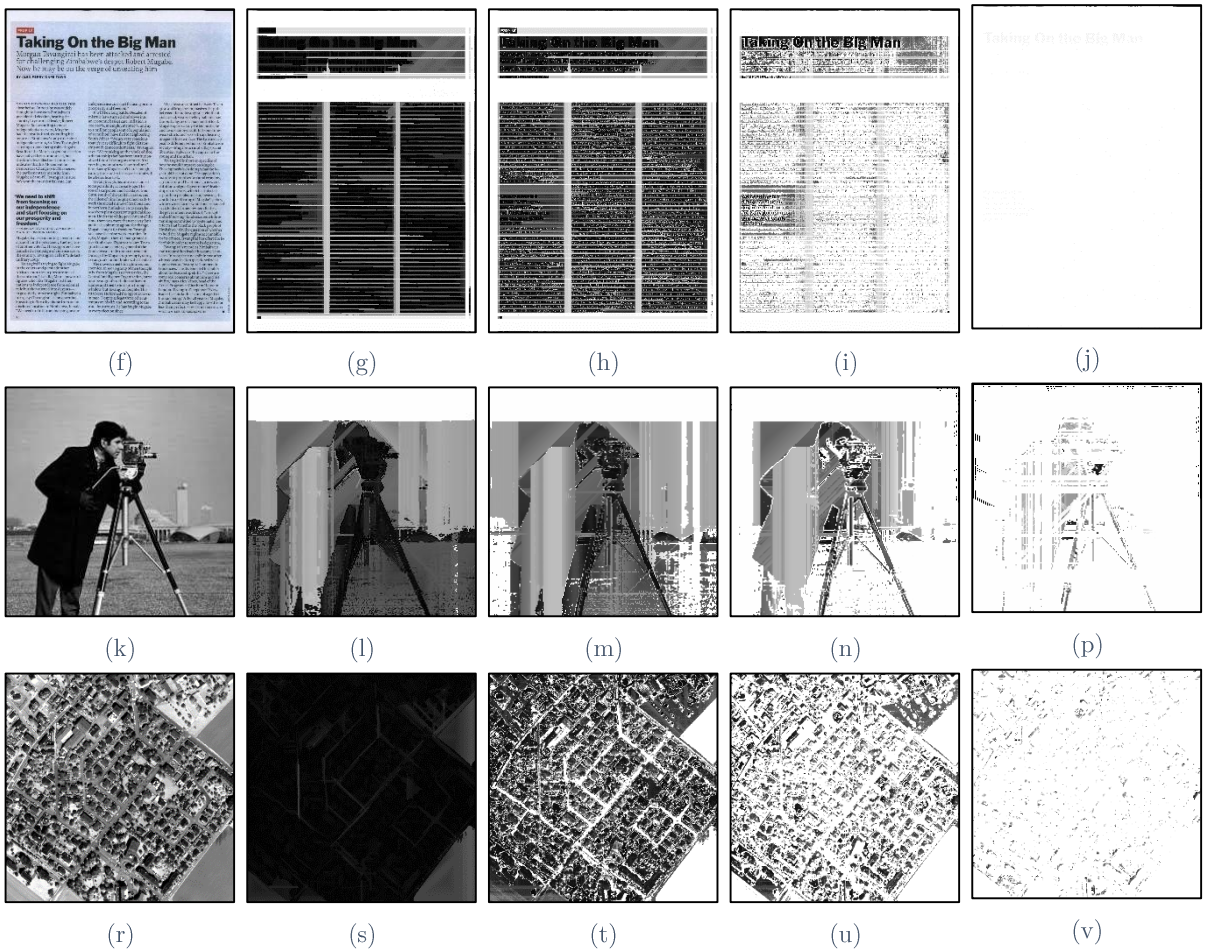
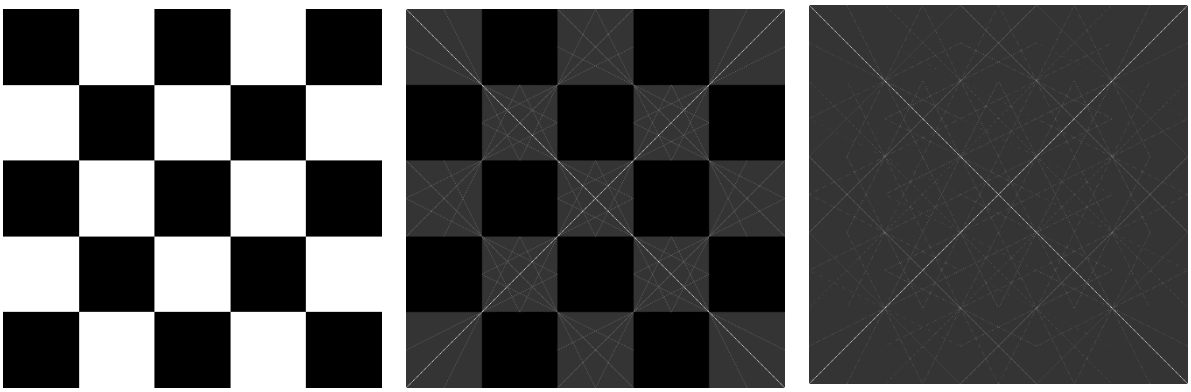


Figure 48 - Applications de le la transformée en diamètre local relative sur des images en niveaux de gris. (a, f, g, r) Image originale, les autres images correspondent à une application de la RLDT en niveaux de gris avec (b, g, l, s : $\beta = 50$; c, h, m, t : $\beta = 100$; d, i, n, u : $\beta = 150$; e, j, p, v : $\beta = 200$).

Une des perspectives serait d'adjoindre à l'image en niveaux de gris, une image binaire même si celle-ci comporte des erreurs. Cela permettrait de donner des indications sur l'endroit où se trouve la forme d'intérêt sans toutefois avoir une confiance totale pour ne pas retomber dans les travers évoqués au début de cette section.

L'autre frein à l'utilisation de la LDT en niveaux de gris est le temps de calcul qui passe de 2N à N^2 pour une direction. En effet, il est impossible d'utiliser l'information calculée sur des pixels déjà traités pour diminuer le temps de calcul global de l'image.



(a) (b) (c)

Figure 49 - Illustration du problème du fond et de la forme. (a) Image originale. (b) RLDT en binaire. (c) RLDT en niveaux de gris ($\beta = 10$)).

2.9 Synthèse et discussions

Nous avons présenté dans ce chapitre quatre transformées dont l'application sur une image binaire, conduit à la définition de différentes caractéristiques possibles (la longueur du segment maximal, l'orientation du segment maximal, *etc.*) et d'un nouvel espace de représentation, celui des traits. De même, ces transformées permettent, en considérant une union de segments de propriétés données, de définir de nouvelles formes, régions de l'image, sur lesquelles des caractéristiques classiques peuvent être calculées.

Une optimisation peut être réalisée en considérant l'enveloppe des composantes connexes. Cela permettrait de modéliser tous les angles sans discrétisation même s'il est à noter qu'elle sera nécessaire pour les transformées en orientation car une infinité de directions pourrait rendre son utilisation moins pertinente. En effet, si nous considérons que chaque segment a une direction différente, nous ne pouvons plus faire de regroupements locaux ou d'analyse. L'un des problèmes majeurs de cette optimisation est celui des traits discontinus entre deux pixels dans l'enveloppe. Si nous considérons tous les couples de points du contour de chaque forme, nous obtenons des « diamètres » qui peuvent être discontinus comme le diamètre sur la droite verte dans la Figure 30 (a).

Le passage aux niveaux de gris peut également être amélioré, notamment sur la notion d'égalité qui définit la forme. De la même manière, nous pouvons peut-être étendre les transformées à des images en couleurs.

Le prochain chapitre sera l'occasion de mettre en œuvre ces transformées. Il présentera l'utilisation de ces dernières et nos contributions pour l'extraction de la mise en page sur des images de documents. Pour ce faire, nous montrerons quelques utilisations possibles de ces transformées appliquées sur deux types de forme : les formes présentes dans le premier plan

dans l'objectif de les caractériser et de les différencier, mais également les formes présentes dans le second plan appelé plus communément le fond, dans l'objectif contraire de les rassembler ou de les délimiter.

Chapitre 3
Extraction de la mise en page dans les documents

Sommaire

Résumé

Dans ce chapitre, nous proposerons une méthodologie pour extraire la mise en page d'un document fondé sur la dualité entre le fond et la forme que nous décrivons en utilisant les transformées définies précédemment. Contrairement aux autres approches qui s'attachent à diversifier les descripteurs, nous privilégions la représentation des éléments à analyser. La primitive segment est à la base de notre proposition. La première étape de notre méthode extrait les séparateurs matérialisés pour en déduire une première partie de la mise en page comme les tableaux. La deuxième étape utilise les séparateurs implicites pour segmenter le document en régions. Enfin, la dernière étape permet de labéliser ces régions. Nous évaluerons les différentes étapes selon le critère de qualité des résultats.

Nous avons abordé, dans le chapitre précédent, différentes transformées permettant de modifier le mode de représentation de l'image. Nous allons voir, dans ce chapitre, comment ces transformées vont permettre d'extraire, dans les images de documents, des informations utiles à la mise en évidence de la mise en page. Nous commencerons par présenter les principes de notre méthode d'analyse de la mise en page (cf. Section 3.1), puis examinerons les différentes étapes pour segmenter (cf. Sections 3.2 et 3.3) et labéliser le document (cf. Section 3.4). Enfin, nous terminerons le chapitre par une conclusion (cf. Section 3.5).

Notations utilisées dans ce chapitre :

Soit I l'image initiale dont le domaine de définition est Δ_f*, en niveaux de gris. Notons* I_{bi} *les images binaires associées à cette image initiale et* $\overline{I_{bi}}$ *leurs images inversées. Les opérateurs* Th^t *et* Th_t *permettent de binariser (seuiller) respectivement les images en niveaux de gris par un seuil t, ne conservant comme formes, que les pixels de niveau de gris respectivement supérieur ou inférieur à* t*.*

3.1 Principe général de la méthode

La mise en page d'un document est un moyen pour celui qui présente un message d'aider à sa lecture et à sa compréhension. Les règles de mise en page évoluent au cours du temps et des usages, mais l'objectif est toujours le même. La mise en page correspond au premier niveau d'observation d'un document et permet de mettre en évidence les différentes parties, ainsi que leur organisation spatiale et structurelle. Nous verrons dans le chapitre suivant (cf. Chapitre 4), dans lequel nous aborderons le problème de la sécurisation d'un document hybride, comment l'extraction de la mise en page physique d'un document peut contribuer à la détection d'un faux ou à l'authentification d'une instance d'un document hybride.

La mise en page comporte différentes couches (tableau, texte, *etc.*) que nous pouvons extraire séquentiellement ou parallèlement. Nous nous sommes intéressés aux différentes couches présentes dans les documents. En analysant les écueils des méthodes actuelles dans le Chapitre 1, notamment celles utilisées par les OCR pour transcrire les documents, nous observons des difficultés dans le double colonnage et les tableaux. En effet, leur présence gêne la segmentation des autres éléments contenus dans les documents. Un tel phénomène est illustré sur la Figure 50, où la méthode de segmentation présente dans l'OCR de Google Tesseract est faussée par la présence de tableaux et de double colonnage. En effet, nous observons dans ces images que des régions labélisées comme lignes sont en fait constituées de corps du texte et d'éléments de tableaux.

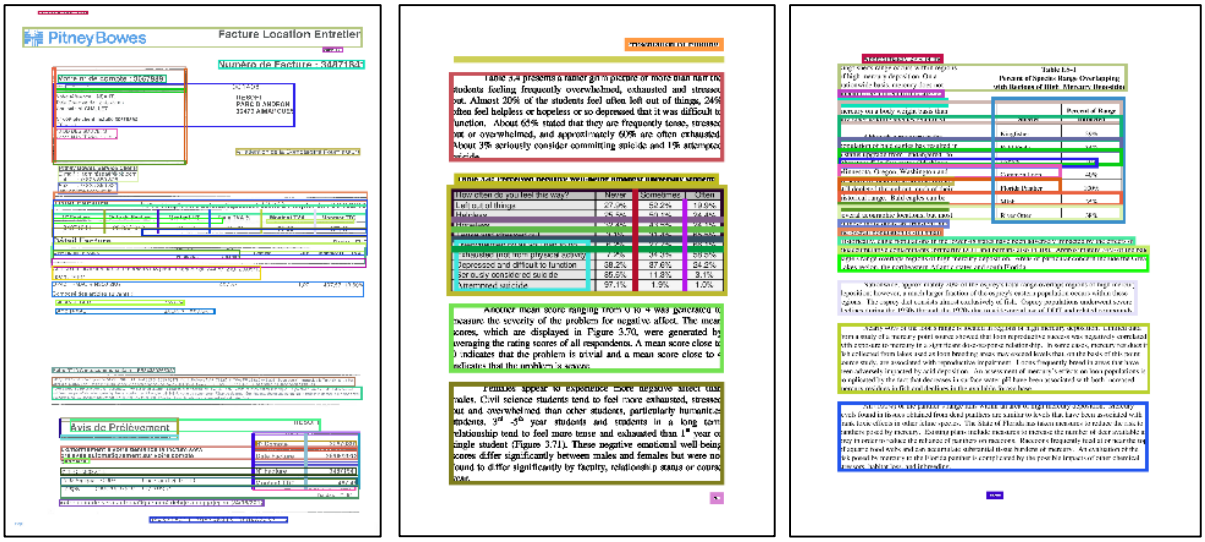
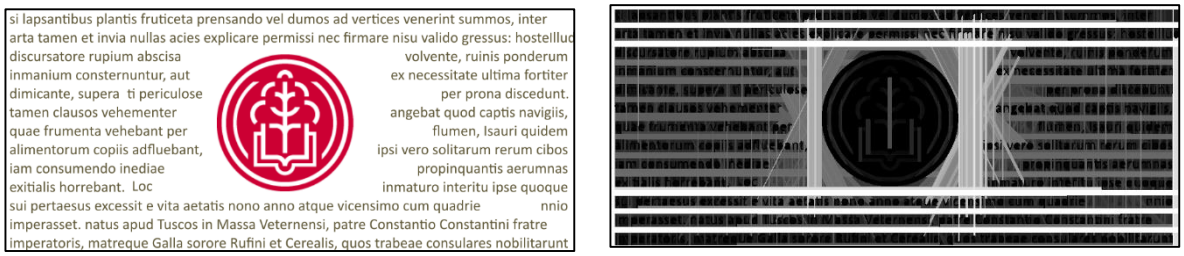


Figure 50 - Exemple d'images de résultats de segmentation de lignes de textes (représentées par des rectangles de couleurs aléatoires) par la méthode Tesseract, ici perturbée par la présence des tableaux et la mise en page en double colonne.

Notre objectif dans cette thèse est d'apporter des solutions à ces différents problèmes. Pour cela, nous adoptons une approche qui traite successivement les éléments les plus évidents dans le document. Puisque la mise en page doit aider notre perception visuelle, nous avons de ce fait analysé les éléments auxquels nous sommes le plus sensibles. Dans le langage courant, on parle de séparateurs. Ils peuvent être explicites ou implicites. Ils sont explicites lorsqu'ils sont matérialisés par un trait plus ou moins épais et plus ou moins long. On les rencontre dans les tableaux, mais aussi pour encadrer une zone textuelle ou illustrative ou séparer deux colonnes. Ils sont implicites quand ils ne sont pas matérialisés. Par exemple, deux colonnes sont séparées par une zone blanche, deux paragraphes sont séparés par un interligne plus grand que deux lignes au sein d'un même paragraphe. En général, ces séparateurs ont une forme relativement linéaire. Les séparateurs implicites dans le fond sont à la base des travaux utilisant les *XY-cut* [MaMS05, Meun05], [NaSe84] ou les rectangles blancs maximums [PhZh91].

Nous avons voulu tirer parti de ces séparateurs et des deux aspects qu'ils peuvent prendre en traitant alternativement la forme et le fond pour extraire des informations complémentaires à partir des transformées définies précédemment. Notre méthode est fondée sur la dualité du fond et de la forme. Des séparateurs sont présents dans les deux parties du document. Les séparateurs de la forme sont explicites tandis que ceux du fond sont implicites. En fonction des mises en page issues de l'imprimerie, cette étude se limitera aux séparateurs linéaires. En effet, même si dans le document le corps du texte encadre un médaillon, les séparateurs circulaires qui l'entourent peuvent être approximés par un ensemble de séparateurs linéaires se raccordant entre eux (cf. Figure 51).



(a) (b)

Figure 51 - Séparateurs implicites présents dans un document comportant un médaillon. (a) Image initiale. (b) Résultat de l'application de la RLDT sur cette image.

Nous avons donc choisi d'extraire du document les séparateurs matérialisés. Ils permettent d'extraire les tableaux entièrement matérialisés. Les traits et les tableaux sont retirés de la forme pour que ceux-ci ne perturbent pas le reste de l'extraction de la mise en page. Une nouvelle définition de la forme est ainsi envisagée. Nous pouvons nous appuyer sur elle pour en extraire la mise en page. Grâce au fond, il est possible de segmenter le document pour regrouper les pixels en régions plus importantes et ainsi mieux les considérer. Les régions ainsi trouvées permettent de calculer des caractéristiques plus fiables sur la forme. Nous avons pu confronter deux méthodes de segmentation pour en améliorer la stabilité face aux documents hybrides. L'étape de segmentation permet d'obtenir des régions qui seront labélisées par la suite en textes et éléments graphiques. Une correction de la segmentation sur les zones de texte, qui sont généralement relativement rectangulaires, permet finalement d'améliorer la stabilité de l'extraction de la mise en page. Le principe général de notre approche est présenté dans la Figure 52.

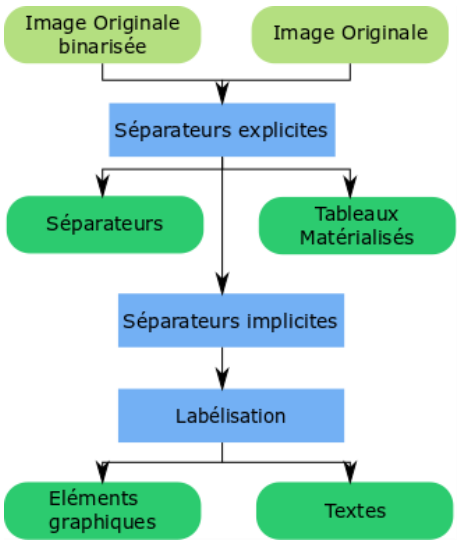


Figure 52 - Organigramme de notre méthode pour l'extraction de la mise en page.

La prochaine section détaille la première partie de notre méthode : le traitement des séparateurs explicites. Cela permet de détecter les tableaux matérialisés et de les retirer pour la suite de l'extraction de la mise en page.

3.2 Extraction des séparateurs explicites

Les séparateurs explicites ont des épaisseurs variables. Certains sont très fins, plus fins que l'épaisseur des caractères. S'ils sont bien visibles sur le document original, leur dégradation, lors de l'acquisition du document numérique, peut être importante. Il est par exemple fréquent, dans les images de documents de voir des traits en partie disparus dans une image binarisée du document. Dans ce contexte nous allons proposer un traitement spécifique adapté à de telles situations.

Notre méthode d'extraction des séparateurs explicites donne lieu à l'extraction de tableaux matérialisés et à la définition de zones de travail plus restreintes que la page (cf. Figure 53).

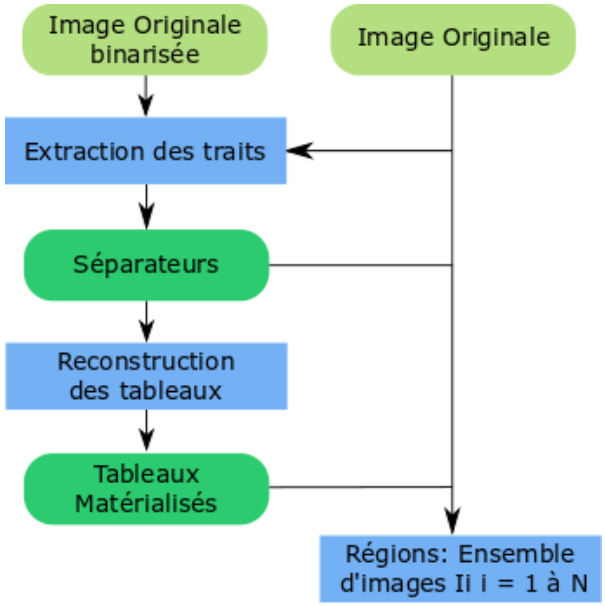


Figure 53 - Organigramme de l'extraction des séparateurs explicites.

3.2.1 Extraction et reconstruction des traits

L'extraction des traits dans une image est une étape délicate du traitement d'images de documents. En effet, lors du passage de la version papier à la version numérique, les traits particulièrement fins peuvent s'atténuer. De manière à préciser la forme dans le document, nous considérons une version binarisée de l'image du document. Pour détecter un séparateur ou plutôt un trait, nous faisons plusieurs hypothèses :

- l'image a été prétraitée pour s'assurer que les lignes de texte sont majoritairement horizontales ;
- les traits ont une trace dans l'image initiale et conservent dans l'image binaire une longueur non négligeable.

La détection est donc opérée en deux temps :

1. sur l'image binaire, les segments de droite sont détectés ;
2. en se référant à l'image initiale en niveaux de gris, les traits sont reconstruits si ceux-ci sont manquants ou détériorés.

➢ **Extraction des traits**

La première étape de notre méthode consiste à extraire les longs segments de forme qui composent l'image du document I. Les séparateurs explicites sont plus facilement détectables à travers une représentation binaire du document. En effet, ceux-ci sont généralement contrastés par rapport au fond du document. La méthode de binarisation est essentielle pour la suite de notre méthode. Il faut que celle-ci soit suffisamment performante pour détecter une part de chaque morceau de trait que nous pourrons rallonger par la suite. Nous avons donc considéré une méthode de binarisation locale permettant de considérer les changements avec plus de finesse. Après expérimentation, nous avons choisi la méthode de Nick [KSFV09] qui permet d'obtenir de bons résultats dans ce cas précis. Dans la même perspective et pour diminuer les erreurs éventuelles, nous avons dilaté la forme dans les images par un élément structurant de taille 3×3. Par abus de notation, nous noterons I_{b1} le résultat de ce traitement.

Nous utilisons la RLDT pour différencier les longs segments, présents dans le document, des autres éléments. Ici les longs segments sont relatifs à la taille du document, ce qui fait que nous utilisons la RLDT et non la LDT. Une analyse préalable nous a permis d'établir que dans un document A4, le corps du texte correspondrait approximativement en moyenne à moins de 2 % de la taille du document. Comme nous nous intéressons aux séparateurs, seuls les éléments les plus grands horizontaux et verticaux nous intéressent. Ainsi, en binarisant l'image issue de la RLDT grâce à ce seuil de 2 %, nous supprimons la plus grande part du bruit. Nous utilisons ensuite la RLOT pour supprimer tous les segments qui ne sont pas verticaux ou horizontaux. Sur la Figure 54, nous pouvons observer le traitement de l'extraction des traits sur un document « difficile », car il a été mal numérisé. La Figure 54 (b) présente la binarisation de celui-ci. Sur la Figure 54 (c), nous observons le résultat de la RLDT appliquée sur la forme de l'image, où les segments courts (moins de 2 %) sont représentés en bleu, les segments de plus de 2 % sont

représentés par les autres couleurs. Nous pouvons observer que le texte apparaît en bleu, de même que des petits morceaux de séparateurs, rendus courts par des problèmes de binarisation. Sur la Figure 54 (d), nous voyons les longs segments qui sont à la fois présents dans les traits des tableaux et dans les cadres de texte (unités de lecture) mais également dans le logo en haut à gauche. Comme les documents que nous considérons sont soit redressés, soit ont une orientation faible, nous ne prenons en compte que les traits horizontaux et verticaux qui sont généralement ceux utilisés dans notre culture moderne comme séparateurs.

Les longs segments sont ainsi définis par la fonction G qui associe à un point $P(x,y)$ d'une image binaire I_{b1} la valeur 0 ou 1 :

$$G(I_{b1})(x,y) = \begin{cases} 1 \; si \; RLDT\,(I_{b1})(x,y) > \; 0{,}02 \; et \; RLOT\,(I_{b1})(x,y) = 0 \; ou \; \pi/2 \\ 0 \; sinon \end{cases} \tag{25}$$

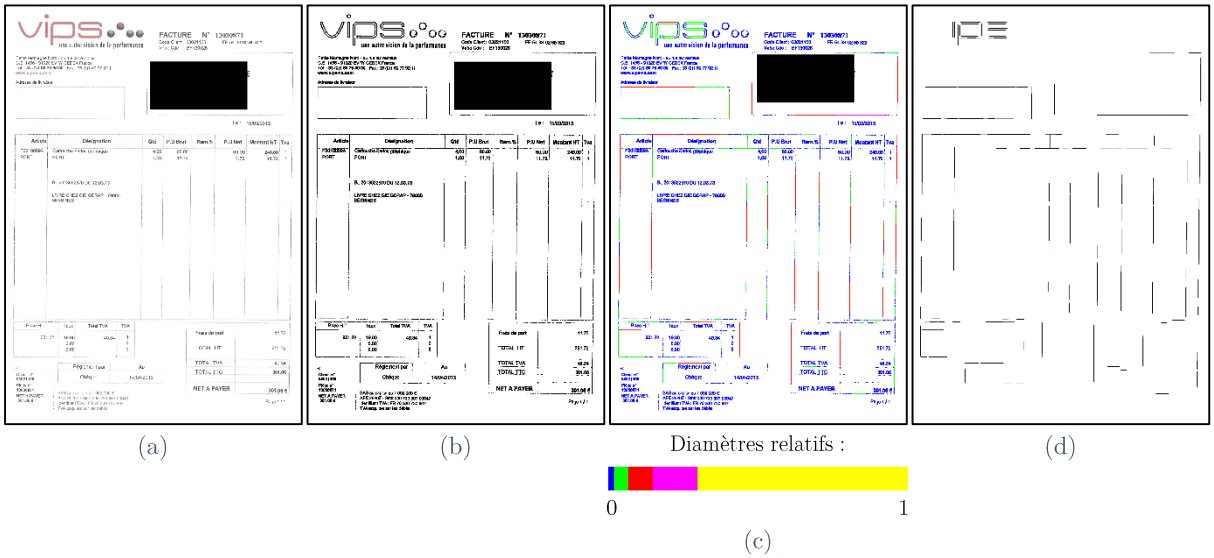


Figure 54 - Extraction des longs segments à partir d'un document difficile car mal numérisé. (a) Image originale I. (b) Image binarisée I_{b1}. (c) RLDT appliquée à l'image binarisée $RLDT(I_{b1})$. (d) Longs segments horizontaux ou verticaux $G(I_{b1})$.

Comme évoqué dans l'introduction, la binarisation n'est pas parfaite pour de nombreuses raisons telles que l'usure ou les variations de niveaux de gris dues à l'acquisition, au scanner, *etc.* De ce fait les longs traits extraits ne le sont pas également. L'une de nos hypothèses de travail est de considérer que la première étape du processus, même si elle présente des points faibles, donne les parties de traits potentiellement effacés. Une nouvelle étape est donc nécessaire afin de compléter et allonger ces segments incomplets que nous désignerons par le terme « germe ».

Dans un premier temps, les germes (cf. Figure 55 (a, b et c)), sont prolongés dans la direction de la RLOT de leurs points jusqu'aux bords de l'image. On obtient alors un

quadrillage $L(I_{b1})$ du domaine Δ_f de l'image (cf. Figure 55 (d, e et f)). Nous faisons alors l'hypothèse que les séparateurs sont présents dans le prolongement à partir du moment où un germe a été détecté dans $G(I_{b1})$.

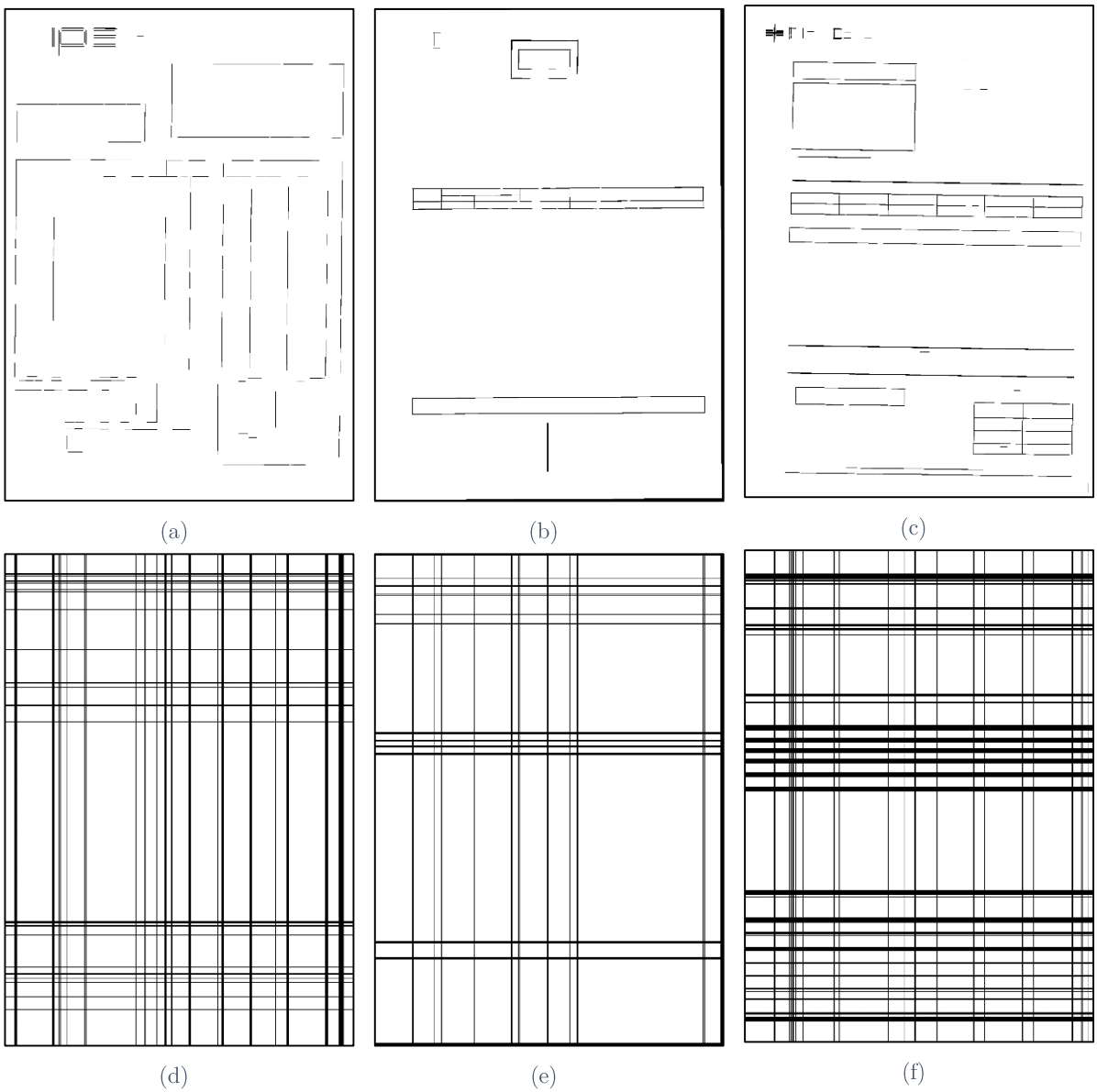


Figure 55 - Exemples des longs segments et de leur prolongement sur trois images de documents. (a, b et c) Résultats de germes $G(I_{b1})$. (d, e et f) Prolongement des segments $L(I_{b1})$.

Les portions de traits ainsi prolongés doivent être vérifiées selon un modèle pour confirmer ou infirmer la présence véritable d'un trait.

➢ Vérification des traits

Si le trait est incomplet, on peut penser que la binarisation n'était pas adaptée à ce contenu. L'information n'est donc pas présente dans I_{b1}. Nous proposons de chercher un

complément d'information dans l'image initiale I. Plutôt que d'extraire le trait et puisque nous avons un indice sur la présence de celui-ci, nous allons procéder à une vérification de sa présence dans l'image initiale en niveaux de gris. Plus précisément, nous proposons un modèle théorique du trait. La présence d'un trait sera confirmée ou infirmée si ce dernier correspond ou non à ce modèle.

Modèle de trait

Le modèle que nous allons utiliser est fondé sur un comportement local d'ensemble de pixels et un comportement global sur la zone de travail. Nous observons l'évolution des niveaux de gris dans I compris entre [0,1] autour d'un trait. La Figure 56, dans sa partie supérieure, représente : à gauche l'évolution des niveaux de gris le long d'une section orthogonale à la direction du trait, à droite, le trait théorique représenté en traits pointillés. Le manque de contraste et les défauts d'impression sur l'image en niveaux de gris se traduisent par la section indiquée en trait plein à droite. Dans la partie inférieure de la Figure 56, le comportement des sections sur la zone de travail est modélisé. Le trait candidat est comparé au comportement théorique sur une zone de travail définie, comprenant de part et d'autre du trait 5 pixels de marge où, de gauche à droite, les niveaux de gris augmentent, se stabilisent (potentiellement) puis diminuent. Sur la Figure 56, la zone rouge correspond à une augmentation des niveaux de gris, la zone verte à une diminution de ces mêmes niveaux de gris et la zone bleue à une stabilité des niveaux de gris le long de la section. Cette variation doit être observée en fonction de la direction considérée. Ainsi, pour la direction verticale, le voisin d'intérêt sera le pixel à gauche tandis que pour la direction horizontale, ce sera le pixel du bas.

Vérification

Dans la zone d'intérêt, nous définissons ainsi trois ensembles :

- Z_i, l'ensemble des pixels dont le niveau de gris augmente par rapport à son voisin d'intérêt (en rouge sur la Figure 56 et la Figure 57) ;
- Z_s, l'ensemble des pixels dont le niveau de gris reste stable par rapport à son voisin d'intérêt (en bleu sur la Figure 56 et la Figure 57) ;
- Z_d, l'ensemble des pixels dont le niveau de gris diminue par rapport à son voisin d'intérêt (en vert sur la Figure 56 et la Figure 57).

Sur un trait hypothétique, la présence de pixels adjacents de Z_i, Z_s et Z_d ou Z_i et Z_d permet de maintenir l'hypothèse d'un trait, là où leur absence confirme la non-présence de

celui-ci. Les éléments de $L(I_{b1})$, dont le modèle présenté ci-dessus est vérifié, constituent l'ensemble de segments $L_r(I_{b1})$.

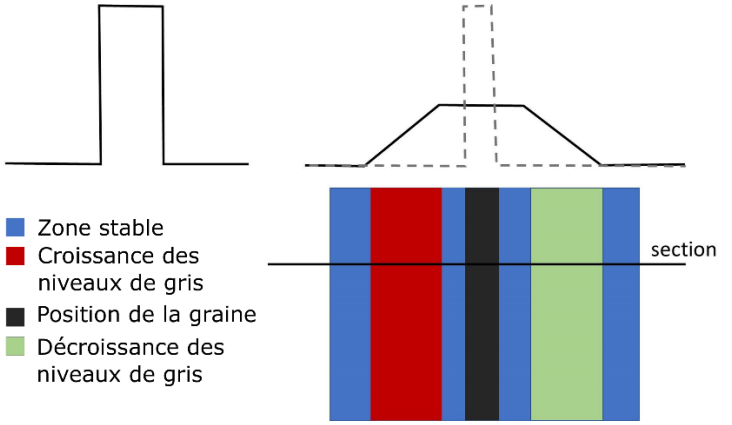


Figure 56 - Modèle de la présence d'un trait.

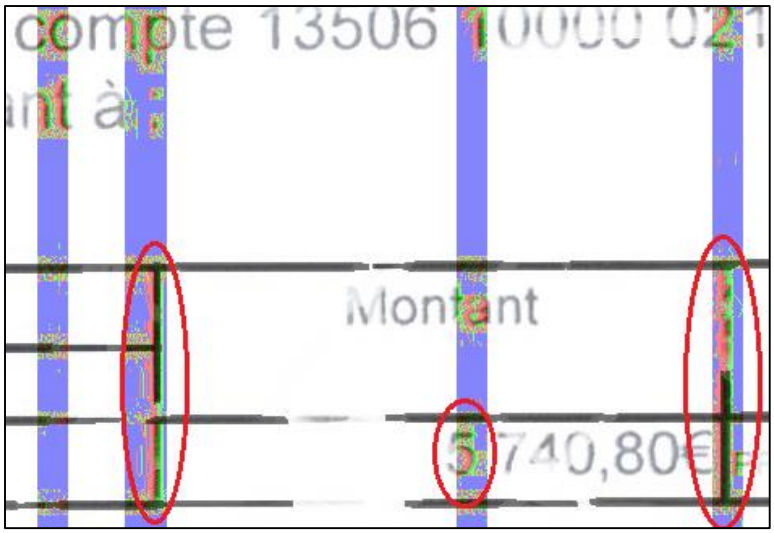


Figure 57 - Exemple du comportement des traits verticaux dans une image dégradée (en bleu : zone stable, en rouge : croissance des niveaux de gris et en vert : décroissance des niveaux de gris).

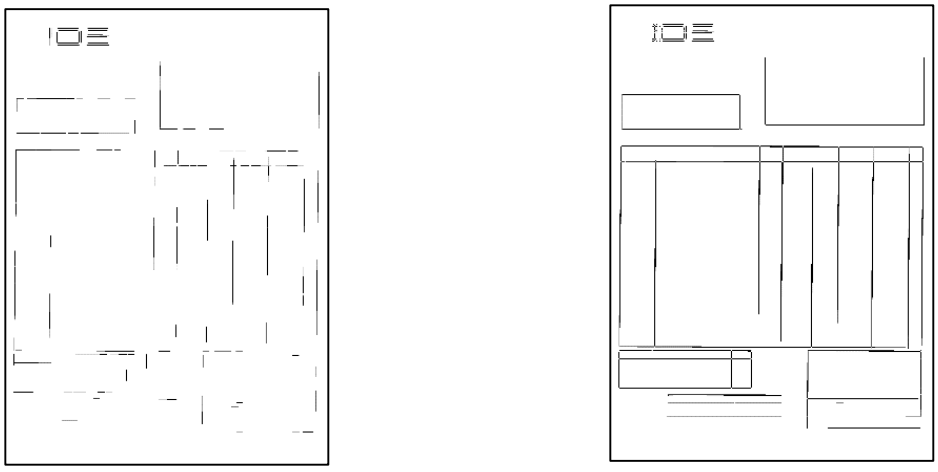
Malheureusement, il arrive que dans l'ensemble $L_r(I_{b1})$, de petits segments soient associés à du texte qui se situe localement dans la même position verticale qu'un segment, comme l'illustre la Figure 57. La dernière étape de la reconstitution des traits consiste à faire la distinction entre les traits isolés et les faux traits associés à du texte.

Les éléments de $L_r(I_{b1})$ sont regroupés selon le quadrillage de $L(I_{b1})$. Soit un élément de $L_r(I_{b1})$ noté s. Dans un vrai trait, $Z_i(s)$ (l'ensemble des éléments de Z_i compris dans s) est un trait dont la longueur est à peu près égale à celle de s. Dans le contexte du texte, la zone $Z_i(s)$ est plus complexe. En effet, celle-ci est constituée de plusieurs petites composantes connexes. La Figure 57 illustre ces deux comportements. Sur les zones entourées en rouge apparaît le cas de vrais traits qui ont disparu et où le comportement du trait correspond au

modèle. Toutefois, sur les zones de texte, nous pouvons également observer une variation locale du rouge vers le vert. Mais, dans ce cas, il n'y a pas une seule grande composante connexe (en 4-connexité) mais une multitude de composantes connexes. Cette différentiation permet de supprimer les composantes connexes considérées à tort comme des traits. Nous pouvons noter que ces remarques sont analogues sur la zone Z_d.

Notons $\{Z_{s,a}\}_{a=1}^{n}$ l'ensemble des composantes connexes en 4-connexité de $Z_i(s)$ et $proj_s(x)$ la projection orthogonale d'un ensemble x sur la direction principale de s. Les indices des éléments de $\{Z_{s,a}\}_{a=1}^{n}$ sont supposés triés dans l'ordre décroissant en fonction de la longueur de leur projection dans la direction de s. Pour accepter l'hypothèse d'un trait, il faut que le nombre de composantes qui suivent le trait ne soit pas trop élevé. Ce nombre dépend évidemment de la longueur du trait. En moyenne, une composante de Z_i pour être significative doit couvrir au moins un dixième de la longueur de s que nous avons nommée n_0. Nous avons donc décidé de ne considérer qu'un nombre n_0 de composantes connexes. Le critère pour considérer un élément de $L_r(I_b)$ comme un trait est que l'union des premières projections des éléments de l'ensemble $\{Z_{s,a}\}_{a=1}^{n}$ ait une longueur supérieure à 90% de la longueur de s, soit :

$$\forall n \leq n_0, \qquad longueur\left(\bigcup_{a=1}^{n_0} proj_s(z_{s,a})\right) > 0.9 * proj_s(s) \tag{26}$$



(a) (b)

Figure 58 - Exemple de résultats de la reconstruction de traits. (a) Image des graines $G(I_{b1})$. (b) Image des traits reconstruits I_s.

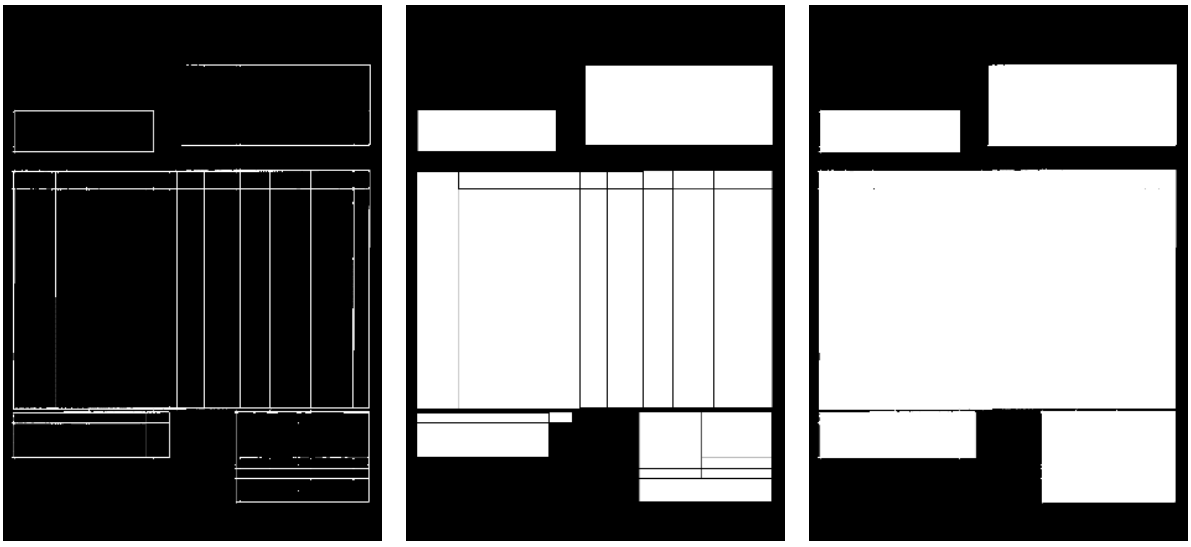
Cette étape permet de supprimer les faux traits. Nous conservons uniquement les segments longs sélectionnés précédemment dans un ensemble final I_s. Des exemples de résultats de l'extraction de traits sont illustrés sur la Figure 58. Ces séparateurs peuvent être isolés ou

constituer des structures plus complexes. Nous présenterons dans la section suivante la détection des tableaux entièrement matérialisés.

3.2.2 Extraction des tableaux matérialisés

L'extraction des tableaux est réalisée à partir des cellules. Une cellule est une composante connexe constituée de traits longs et qui ne contient pas d'autres composantes connexes. Les composantes connexes de I_s peuvent être adjacentes à des caractères de texte si ceux-ci touchent un trait. Soit I^c l'image complémentaire de l'image binaire I_b. L'image I^c représente donc le fond du document. Ce complémentaire nous permet de trouver les tableaux potentiels (cf. Figure 59 (c)).

Les composantes connexes de I^c non bornées ne peuvent pas faire partie d'un tableau et constituent les régions R dans le domaine de l'image. Notons le complément de R, R^c. Celui-ci comprend les pixels des lignes S et les pixels du fond B (cf. Figure 59 (b)). Les composantes connexes de R^c sont étiquetées comme des tableaux si elles contiennent des pixels de B (donc des pixels appartenant à des cellules potentielles) (cf. Figure 59). Sinon, elles correspondent à des lignes de séparation isolées.



(a) (b) (c)

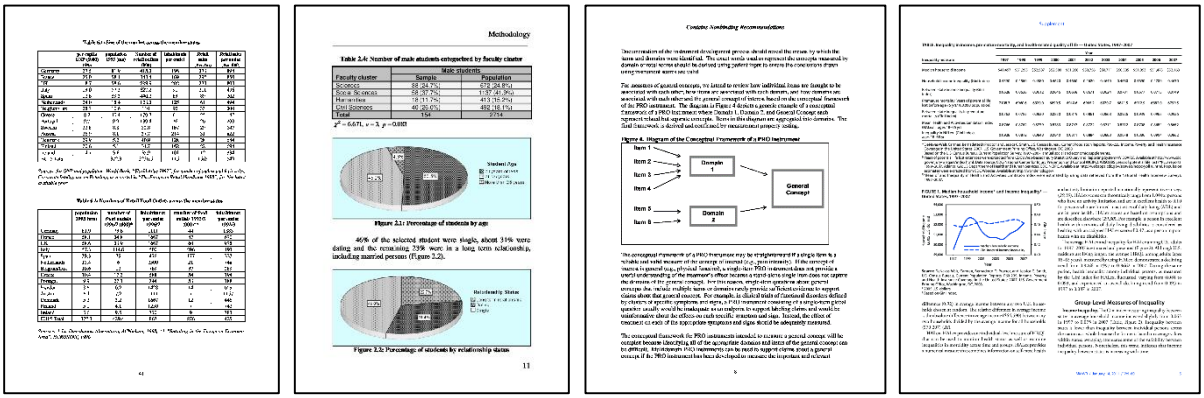
Figure 59 - Illustration sur un exemple des étapes de l'extraction des tableaux. (a) Traits reconstitués (I_s). (b) Cellules potentielles. (c) Tableaux potentiels.

- **Évaluation expérimentale de l'extraction des tableaux**

Nous avons évalué l'intérêt de cette méthode *via* la qualité des résultats obtenus sur deux jeux de données. Le premier résulte de la compétition ICDAR de 2013 sur l'extraction des tableaux [GHOO13] et le second est une base d'images de documents fournie par notre partenaire ITESOFT nommée « SETSTABLE dataset », comprenant des instances de

documents hybrides. Il nous a paru intéressant d'évaluer les résultats sur ces deux bases car la première comportait des documents « propres », sans dégradation mais ne contenait pas plusieurs instances d'un même document hybride ; la seconde comportait, quant à elle, des documents scannés par notre partenaire industriel dans un contexte d'exploitation industrielle, avec plusieurs instances d'un document hybride.

La compétition d'ICDAR [GHOO13] proposait comme tâche d'extraire automatiquement les tableaux dans les documents au format PDF (cf. Figure 60). Les tableaux présents dans la base utilisée sont un mélange de tableaux entièrement matérialisés, ou non, avec des pages comportant des tableaux et d'autres non. La méthode présentée ne permet pas d'extraire les tableaux qui ne sont pas entièrement matérialisés. Nous avons donc créé une sous-base en éliminant toutes les pages comportant ces tableaux. Par ailleurs, notre étude portant également sur des images de documents, nous avons rasterisé les PDF avec une résolution de 150 dpi et 300 dpi, avec des résultats identiques dans les deux résolutions. La sous-base contient 179 images de documents dont 69 contiennent des tableaux.



(a) (b) (c) (d)

Figure 60 - Exemples d'images de la base issue de la compétition d'extraction de tableaux d'ICDAR 2013 [GHOO13].

Pour évaluer quantitativement la qualité des résultats fournis par notre approche, nous considérons comme :

- vrai positif (VP) un tableau T extrait dont l'aire de la surface superposée avec celle du tableau de la vérité terrain T_{GT} est supérieure ou égale à 80 % de l'union des deux aires soit $\frac{A(T \cap T_{GT})}{A(T \cup T_{GT})} \geq 0{,}8$;
- faux négatif (FN) un tableau T existant dans la vérité terrain mais qui n'a pas été détecté ou dont l'aire de la surface superposée avec celle du tableau de la vérité

terrain T_{GT} est inférieure à 80 % de l'union des deux aires soit $\frac{A(T \cap T_{GT})}{A(T \cup T_{GT})} < 0{,}8$;

- faux positif (FP) un tableau extrait mais qui n'est pas présent dans la vérité terrain.

Cela permet de calculer la précision et le rappel :

$$précision = \frac{VP}{VP + FP} \quad ((27)$$

$$rappel = \frac{VP}{VP + FN} \quad ((28)$$

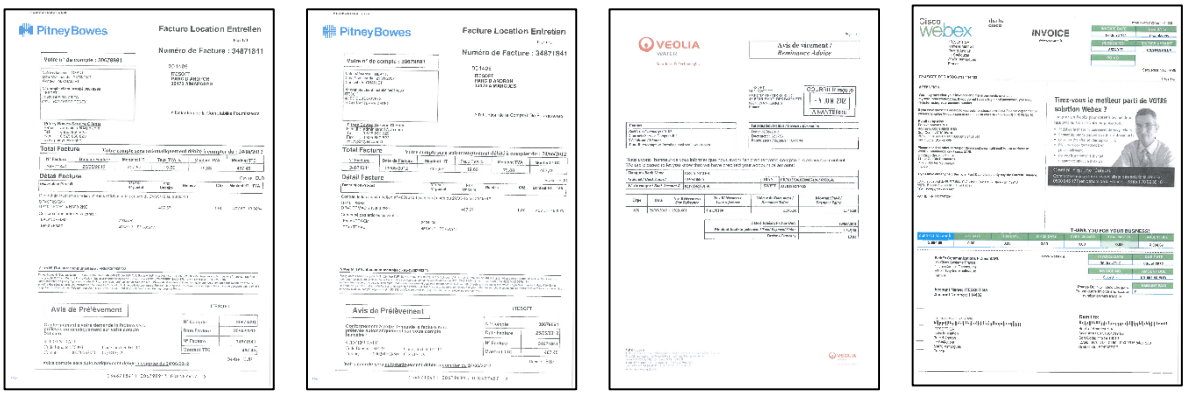
Le Tableau 4 présente les résultats obtenus sur la sous-base d'ICDAR 2013, en les comparant à ceux extraits grâce à la méthode de Tesseract présentée dans [ShSm10]. Pour cette méthode, nous avons utilisé l'implémentation open source de cet algorithme fournie en même temps que l'OCR Tesseract avec les valeurs de paramètres recommandées. Cependant, nous ne pouvons pas nous comparer avec l'ensemble des méthodes ayant participé à la compétition car leurs résultats étaient relatifs à toute la base et non à la sous-base considérée. La vérité terrain contient 88 tableaux et notre méthode en a correctement identifié 86 et en a détecté 18 supplémentaires. Ces tableaux supplémentaires sont souvent le résultat de traits se croisant dans d'autres éléments du document comme le montre la Figure 62 (b), où nous avons détecté un tableau dans un histogramme. Nos résultats sont meilleurs que ceux de Tesseract.

Tableau 4 - Résultats de qualité sur la sous-base d'ICDAR 2013 [GHOO13].

Méthode	Précision	Rappel	F-mesure
Tesseract [ShSm10]	0,25	0,33	0,28
Méthode proposée	0,98	0,83	0,90

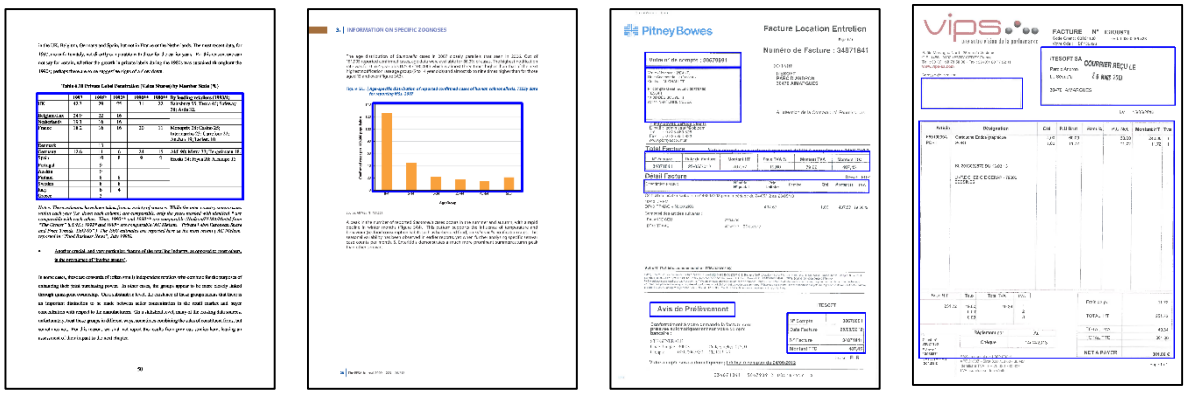
Le premier jeu de données ne contient que des images « propres » où la reconstruction de ligne n'a pas de réel intérêt. Nous avons donc choisi de présenter les résultats sur une autre base « SETSTABLE dataset » où 14 documents ont été scannés plusieurs fois. Cette base contient 293 images de documents scannés à 300 dpi (cf. Figure 61). Nous pouvons ainsi comparer les résultats sans l'étape de reconstruction des traits (LRS) et avec celle-ci pour évaluer son intérêt. Le Tableau 5 montre que sans l'étape de reconstruction des traits, nous avons un score de F-mesure de la détection des tableaux relativement faible (0,54), cela s'explique par l'état dégradé des images de documents. Mais lorsque l'on utilise l'étape de reconstruction des traits, nous augmentons la précision de presque 10 % et le rappel de 20 %. Cette évaluation montre l'efficacité de la reconstruction des traits qui permet d'améliorer

significativement la précision et le rappel. En effet, de nombreux tableaux ont pu être détectés là où les lignes étaient très dégradées.



(a) (b) (c) (d)

Figure 61 - Exemples d'images de la base « SETSTABLE dataset ».



(a) (b) (c) (d)

Figure 62 - Exemples de résultats de notre méthode pour l'extraction de tableaux. (a et b) Exemples d'images résultats obtenus sur la sous-base d'ICDAR 2013 [GHOO13]. *(c et d) Exemples d'images résultats obtenus sur le jeu de données « SETSTABLE dataset ».*

Tableau 5 - Résultats de la qualité sur « SETSTABLE dataset ».

Méthode	Précision	Rappel	F-mesure
Notre méthode sans LRS	0,64	0,46	0,54
Notre méthode avec LRS	0,73	0,66	0,70

Les tableaux ne sont pas les seuls éléments que nous pouvons déduire des segments, ils permettent également de définir des zones de travail comme nous allons le voir dans la section suivante.

3.2.3 Définition de zones de travail

De manière à réduire la complexité du processus d'extraction de la mise en page nous utilisons les séparateurs matérialisés extraits précédemment pour isoler des zones de l'image

qui ne peuvent pas se trouver dans la même entité de segmentation de la mise en page. À partir de la détection des tableaux chaque cellule constitue une zone de travail.

On dispose ainsi d'un ensemble de régions R_i dans l'image auxquelles on va associer autant de nouvelles images traitées par la suite. Nous définissons I^i, l'intersection de la région R_i et de l'image originale I.

Après avoir extrait les tableaux et les traits du document, nous recherchons les autres zones de travail dans le reste du document. Pour ce faire, nous modifions notre image de travail I pour remplacer les régions R_i par du fond. Nous pouvons ensuite segmenter l'image du document pour en extraire la mise en page.

3.3 Segmentation par les séparateurs implicites

Le fond du document comporte de multiples informations. En effet, c'est par ce moyen que nous percevons les différentes parties du document. L'espace compris entre les différents éléments permet à l'œil humain de structurer les informations et de comprendre le sens de lecture. Partant de ce constat, nous proposons d'approximer le document et donc sa mise en page par le contenu du fond du document. La distinction de la forme et du fond est une hypothèse forte sur laquelle nous fondons notre méthode. La méthode de binarisation considérée, dans ce cas-ci, doit être sensible aux faux négatifs (éléments considérés comme appartenant au fond alors qu'ils appartiennent à la forme) quitte à avoir plus de faux positifs (éléments considérés comme appartenant à la forme alors qu'ils appartiennent au fond). Nous avons choisi de combiner les résultats d'une binarisation globale (Otsu [Otsu79]), qui donne une bonne approximation du seuil global optimal et la binarisation locale que nous avions précédemment utilisé (Nick [KSFV09]). Nous notons I_{b2} la composée de ces deux binarisations. La segmentation suivante peut s'appliquer à chaque image I^i que nous avons définie précédemment.

3.3.1 Principe

Le principe de l'approche proposée de segmentation par le fond du document est d'extraire les grands segments présents dans le fond qui correspondent aux séparateurs implicites. Mais lorsque l'on considère uniquement les segments contenus dans le fond, et la grandeur relative de ceux-ci par rapport au document, nous observons que plus le document est « vide », moins l'importance des segments n'a de sens. Ainsi un document ne contenant qu'une phrase aura des séparateurs implicites entre chaque caractère (cf. Figure 63).

Limiter les segments considérés revient à faire une approximation plus ou moins précise de cette partie imprimée. Pour trouver les séparateurs implicites présents dans le fond, il convient de calculer la RLDT du fond du document notée $RLDT(\overline{I_{b2}})$ (cf. Figure 66 (c)). Une première étape consiste à extraire du fond les grandes zones permettant de restreindre les zones de travail (cf. Section 3.3.1) que nous segmenterons ensuite (cf. Section 3.3.2) (cf. Figure 64).

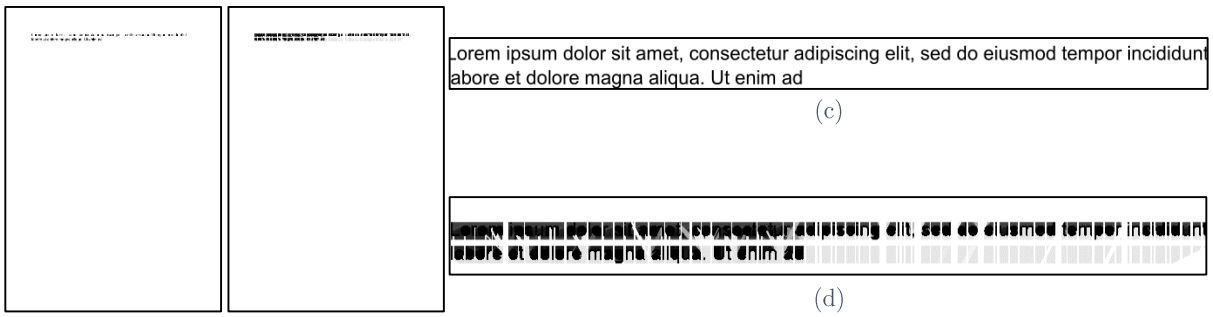


Figure 63 - Illustration présentant le problème de la segmentation par les séparateurs implicites dans les documents « vides ». (a) Image initiale ne contenant qu'une ligne de texte. (b) RLDT appliquée sur le fond de l'image initiale (a). (c) Zoom de l'image représentée en (a). (d) Zoom de l'image représentée en (b).

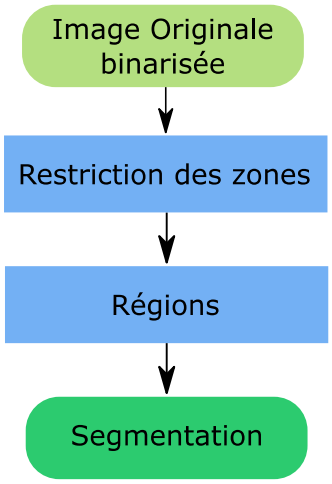


Figure 64 - Organigramme de la segmentation.

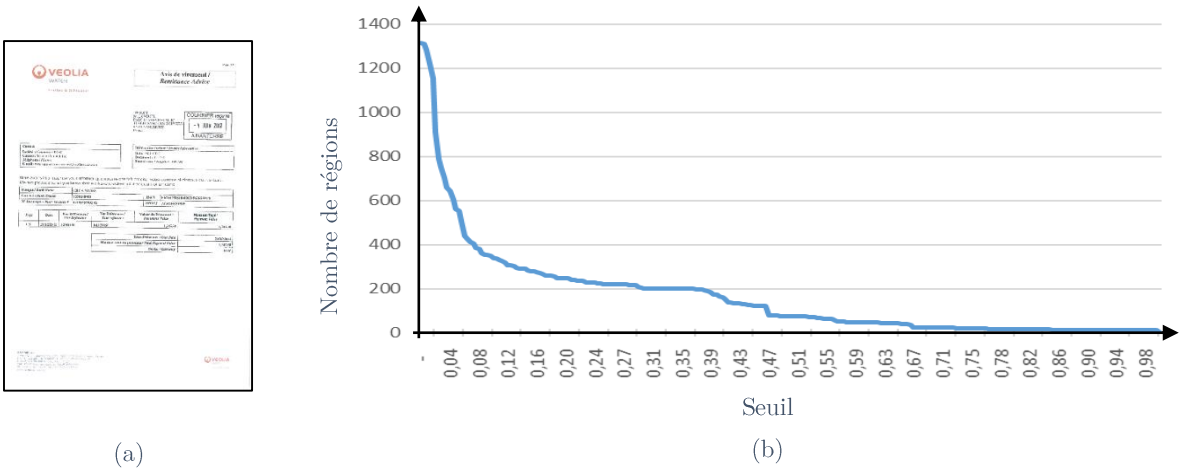
3.3.1 Restriction par grandes zones

En considérant les grands traits implicites, ceux qui traversent presque entièrement le fond du document, nous pouvons améliorer la qualité de la segmentation. La segmentation en considérant uniquement les grands traits dans le document, peut être mise en parallèle avec les premières étapes de la segmentation X-Y cut [NaSe84]. Celle-ci segmente itérativement un document lorsqu'un trait de fond le parcourt entièrement.

Les segments associés aux pixels entre les caractères sont beaucoup plus courts, quelles que soient leurs directions, que les segments associés aux pixels entre les lignes. Ceux-ci sont le

plus souvent horizontaux. Ainsi, plus nous considérerons des segments courts, plus une segmentation au niveau « caractère » sera obtenue. Plus on se restreindra aux grands segments, moins la segmentation sera fine et plus elle sera grossière.

Pour illustrer l'effet de la valeur du seuil dans l'approximation des zones d'intérêt du document, nous avons fait varier la valeur et compté le nombre de régions construites. Le résultat est illustré sur le graphe de la Figure 65. On peut observer, à gauche, un grand nombre de zones qui correspondent aux caractères, puis une forte diminution de ce nombre quand les caractères d'une ligne ne sont plus séparés et constituent une ligne à part entière. Les brusques changements dans cette courbe correspondent à la fusion de régions proches.



(a) (b)

Figure 65 - Influence du seuil utilisé pour segmenter l'image de la RLDT du fond sur le nombre de régions de l'image (a).

Restreindre les zones permet d'éviter les problèmes du fond que nous rencontrons dans les documents ayant peu de contenu, comme nous le verrons dans le chapitre suivant. Cette pré-étape de segmentation nous permet de créer des zones de travail qui sont les composantes connexes de l'image extraite dans $TH_{99}\,(RLDT(\overline{I_{b2}})$ (cf. Figure 66 (f) et (i)). Comme nous l'avons défini dans la Section 3.2.3, chaque région permet d'extraire une image I^i. Les régions obtenues sont encore divisibles. Nous allons donc les diviser grâce à notre méthode de segmentation utilisant également les segments du fond.

3.3.2 Stratégie de segmentation

Dans cette section nous présentons comment segmenter chaque image I^i. Expérimentalement, nous avons observé que considérer les segments dont la longueur est supérieure à 7 % de la taille du document conduisait à trouver l'enveloppe des lignes de texte et des éléments graphiques (cf. Figure 66 (e)). Les composantes connexes du complémentaire de cet ensemble de segments forment les régions. Celles-ci seront, par la suite, labélisées. En

ne considérant que les segments de longueur supérieure à un certain seuil, nous pouvons créer artificiellement des régions qui n'ont pas de réelle signification car elles sont vides (cf. Figure 67). Notons $SW(I^i)$, l'image de segmentation contenant les composantes connexes de $Th^{0.07}(RLDT(\bar{I}^i))$ dont le contenu dans I_{b2} est non vide. On notera R l'ensemble de ces composantes connexes.

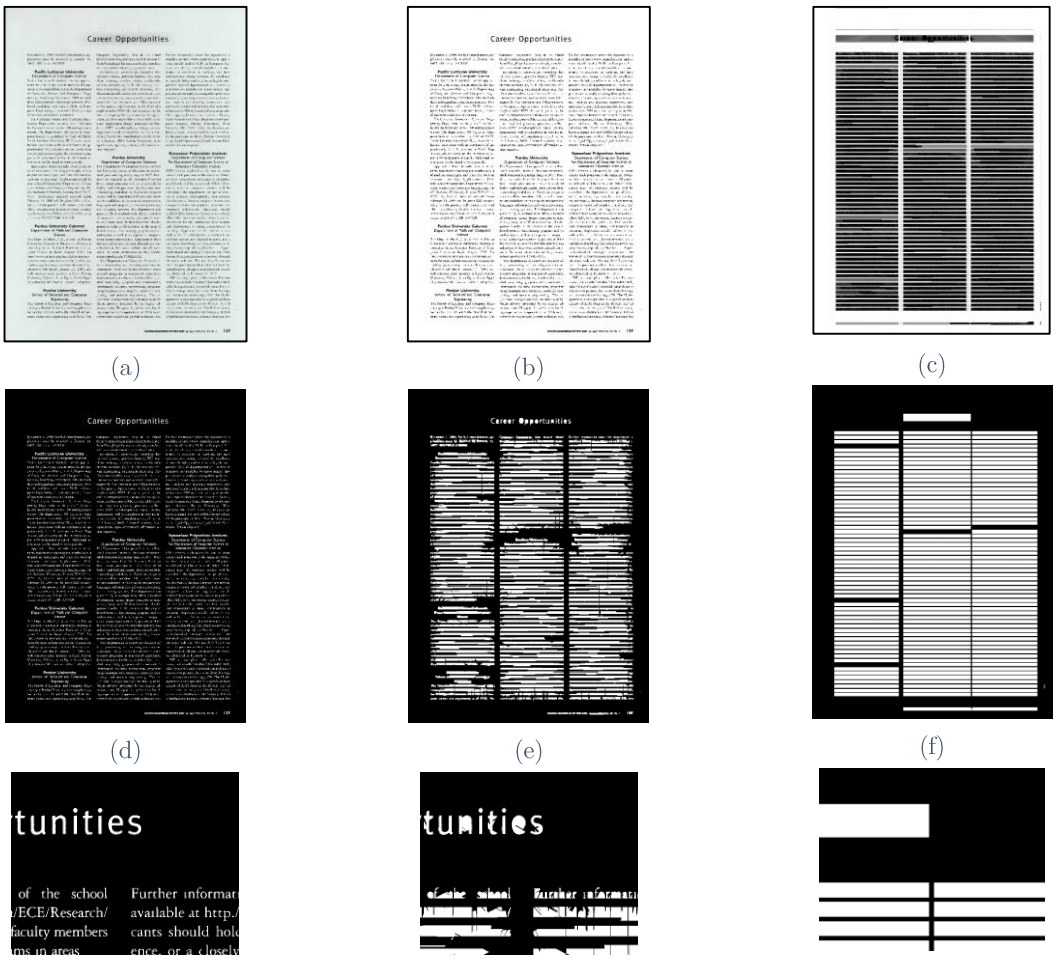


Figure 66 – Illustration des longs segments du fond. (a) Image originale. (b) Image binaire. (c) Image de la RLDT du fond. (d) $Th^0(RLDT(\bar{I}))$. (e) $Th^{0.07}(RLDT(\bar{I}))$. (f) $Th^{0.99}(RLDT(\bar{I}))$. (g, h et i) Zoom des images d, e et f.

Figure 67 - Illustration du problème des régions vides. (a) RLDT sur le fond de l'image contenant le mot « Contents » (échelle continue du rouge égale à 1 au bleu proche de 0. (b) Binarisation avec un seuil fixé à 0.07 de (a) (avec en vert les composantes connexes vides par rapport à l'image binaire).

Si nous avons ici considéré les segments comme des primitives, nous allons désormais les utiliser pour définir des caractéristiques associées aux zones à labéliser et considérer successivement les points de vue qui nous semblent caractériser les propriétés des séparateurs, du texte et des images.

3.4 Labélisation

Les étapes précédentes ont permis d'obtenir différents éléments qui pourront être combinés pour extraire la mise en page des documents. Il est ainsi possible de labéliser les régions précédemment trouvées.

3.4.1 Labélisation de zones de texte et remise en cause de leurs contours

Nous nous intéressons à la labélisation de deux types de textes différents : les titres en « vidéo inverse » (texte clair sur fond foncé) et le corps du texte généralement foncé sur un fond clair. Dans un premier temps, les titres dont la taille est très importante (comme on peut en trouver dans les magazines) ne seront pas traités.

Nous commencerons par labéliser les zones de texte grâce à l'utilisation de la forme combinant l'information des régions précédemment extraites, puis remettrons en question leurs enveloppes, car le texte est généralement un élément rectangulaire.

➢ Labélisation des zones de texte

Pour labéliser les zones de texte, deux points de vue complémentaires seront considérés simultanément. D'une part, un aspect d'évidence par approximation comme il a été fait pour la détection du fond et, d'autre part, le point de vue des régions créées précédemment.

Les zones de corps du texte peuvent être caractérisées par les traits courts qui les composent dans le contenu de l'image et qui définissent les caractères. Nous avons observé

expérimentalement que, considérer les segments dont la longueur est inférieure à 2 % de la taille du document, permettait d'approximer la plupart des caractères du corps du texte du document. Dans une région $r \in R$ à labéliser, la réunion de ses segments courts K est utilisée comme un marqueur indiquant une possibilité de présence de texte. Si cette région r est une région de texte, K devrait être une bonne approximation de la restriction à r de l'image binaire notée Γ_r. C'est-à-dire que la région devrait être bien recouverte par les petits segments. C'est la qualité de l'approximation que nous considérons pour prendre une décision. Nous définissons un degré de confiance doC vis-à-vis d'un label texte calculé pour chaque composante de R. Il peut être alors mesuré comme le pourcentage de l'aire de R par rapport à Γ_r de la façon suivante :

$$doC_K(r) = \frac{Aire\ (\Gamma_r \cap K)}{Aire\ (\Gamma_r)}$$

L'application de cette fonction sur toutes les régions de R se note : $doC_K(I)$. Sa valeur sur le fond étant fixée à 0, les composantes connexes de $SW(I)$ labélisées comme texte sont alors définies par $Te(I) = Th^t(doC_K(I))$, où, expérimentalement, une valeur de t égale à 0,85 a été choisie.

Nous proposons de détecter le texte en « vidéo inverse » par la caractérisation du « fond du texte . Dans l'image I, lors de la binarisation, les interlignes de textes apparaissent reliées par des inter-caractères. La région est alors caractérisée, dans un espace de dimension 1, par le nombre de directions différentes des segments longs présents dans la région. Une région qui n'a pas été labélisée « séparateur » et qui ne comporte qu'une seule direction de segments longs est alors considérée comme région de texte en « vidéo inverse ».

Les titres en « vidéo inverse » sont détectés en analysant les grands traits présents dans le contenu de l'image. Pour les étudier, nous nous plaçons dans un espace de représentation induit par les caractéristiques relatives à l'orientation des segments, en calculant $RLOT(I_{b2})$. Les grands segments sont extraits grâce à la RLDT appliquée sur l'image binaire I_{b2}. Après avoir extrait des régions R toutes les régions précédemment labélisées, nous examinons si les éléments de $R \cap Th^{10}\ RLDT(I_{b2})$ possèdent la même direction. S'ils respectent cette condition, ils sont labélisés comme texte et retirés de R.

- **Remise en question de la segmentation sur les régions de texte**

Notre approche de segmentation est une segmentation orientée « lignes ». Or, plus notre segmentation est fine, plus les régions sont nombreuses, et, plus la stabilité pour les documents hybrides est difficile à obtenir. Nous avons donc créé une méthode permettant de passer de la

segmentation en lignes à une segmentation en blocs. Nous pouvons observer que les régions de textes sont des éléments particuliers qui, dans les documents contemporains que nous sommes amenés à traiter, sont généralement de forme rectangulaire. Notre segmentation est sensible aux cheminées à l'intérieur d'un paragraphe. Nous entendons par là, des espacements alignés dans le texte formant de longs séparateurs blancs de fond (cf. Figure 68). La segmentation est également sensible aux premières et aux dernières lignes proches du fond (cf. Figure 66 (h)).

Pour améliorer la segmentation des zones de texte, nous considérons les régions de texte par leurs boîtes englobantes. Deux critères peuvent nous permettre de fusionner deux régions : leurs espaces inter boîtes et leurs alignements communs.

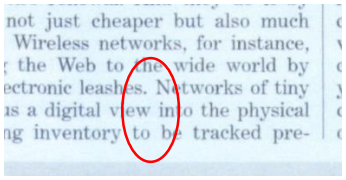


Figure 68 - Illustration des cheminées dans le texte pouvant gêner sa segmentation.

Comme nous l'avons évoqué précédemment, certaines régions de texte ont été trop divisées (sur-segmentées). Par étude expérimentale, nous savons que des mots appartenant au même paragraphe ont un espace inter-mots ou inter-caractères inférieur à deux fois leur hauteur. Nous avons donc pu, dans un premier temps, fusionner toutes les lignes alignées dont la proximité verticale respectait cette contrainte.

Sur les documents, et pour faciliter la lecture nous alignons généralement les lignes de texte faisant partie du même paragraphe que ce soit sur la droite, la gauche ou le centre. Nous fusionnons ces lignes de texte, en comparant l'alignement de deux régions consécutives dont la proximité est inférieure à deux fois la hauteur du plus petit, quand celles-ci sont alignées.

3.4.2 Labélisation des éléments graphiques

Pour continuer la labélisation, il nous faut changer de point de vue. En effet les régions non encore labélisées peuvent être aussi bien du texte ayant une taille non adaptée au choix du seuil de 2 % considéré précédemment que des éléments graphiques.

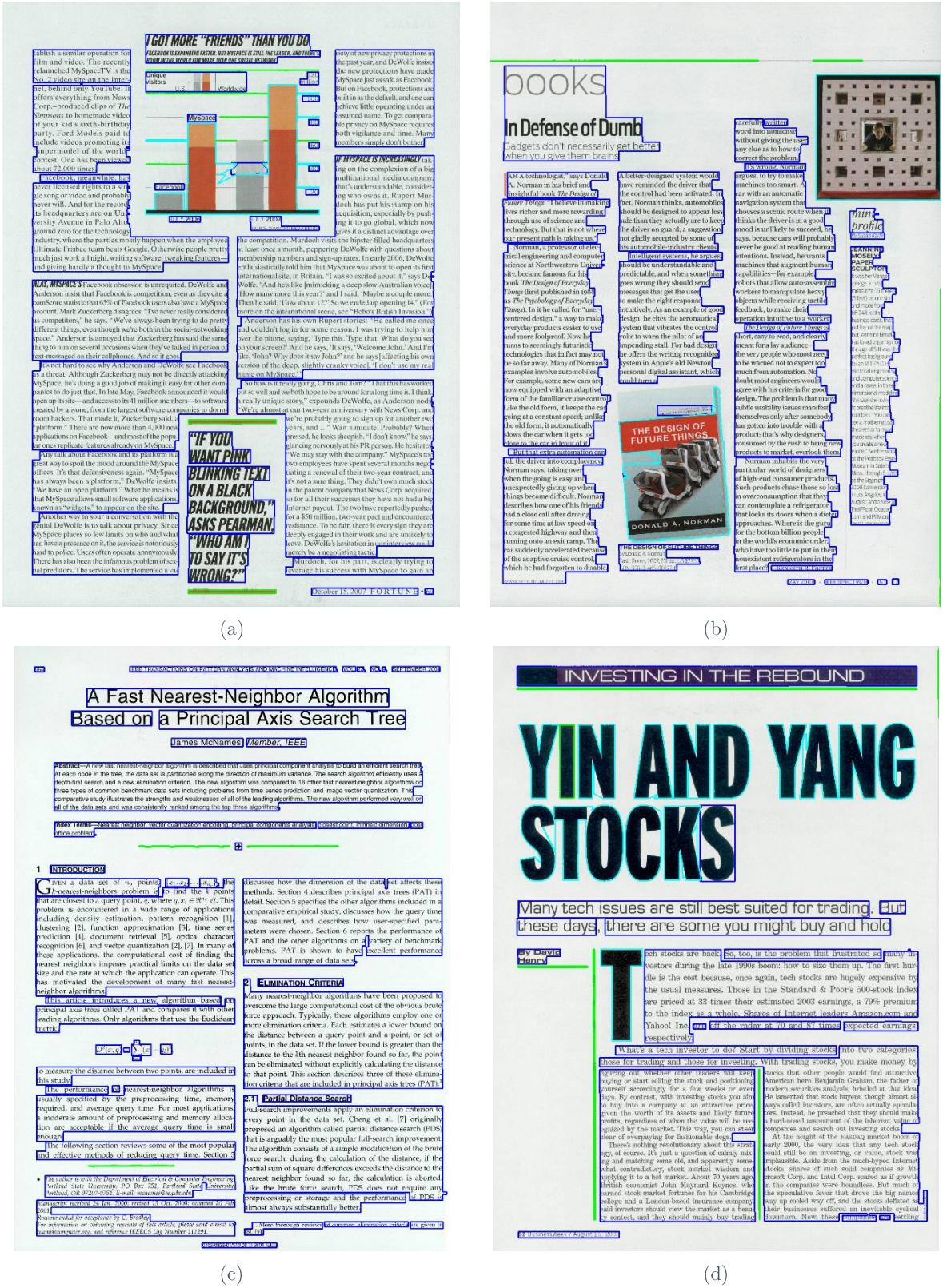
Ce sont les caractéristiques couleur de l'image initiale I qui nous permettent de distinguer un élément graphique, qui comporte plusieurs couleurs, d'un texte que nous avons considéré comme étant de couleur uniforme. Nous nous plaçons donc, pour caractériser une région, dans un espace de dimension 1 où la caractéristique considérée est le nombre de couleurs de la région.

3.4.3 Évaluation de l'extraction de mise en page

Dans un premier temps, nous avons évalué la méthode d'extraction de la mise en page sans les étapes d'extraction des séparateurs implicites (cf. Section 3.2) et de restriction des zones (cf. section 3.3.1). Cette méthode a fait l'objet d'un article en 2017 dans la conférence ICDAR [ACKO17]. Dans la suite de cette section nous la nommerons « Méthode proposée (2017) ». À des fins d'étude comparative, nous avons confronté notre méthode à celle de Fehli *et al.* [FeTS14], évaluée sur la base PRIma [APBP09]. Cette base est issue d'une compétition présentée en 2009 portant sur l'extraction de la mise en page de documents. Cette compétition a lieu à chaque édition d'ICDAR mais en 2009 ce fut la dernière édition où la différenciation des labels dans la présentation des résultats était proposée. La base de données contient 55 images issues de magazines ou d'articles scannés en 300 dpi, qui constituent une sous partie de la base PRIma (cf. Figure 3).

Les résultats obtenus par notre méthode sont évalués et comparés à ceux des méthodes de la compétition présentés dans le Tableau 6. Nos résultats sont meilleurs dans les deux premiers labels, image et séparateur, et légèrement moins bons dans le cadre de la détection de texte car notre méthode conduit à une sur-segmentation des paragraphes. Avec notre nouvelle méthode, nous obtenons une amélioration sur le texte. Cela s'explique par le fait que les lignes sont moins isolées que dans les autres cas. Nous observons sur la Figure 69 les résultats de l'extraction de la mise en page avec en vert les séparateurs, en bleu foncé le texte et en bleu clair les images. Les erreurs présentes sur ces documents sont des erreurs typiques que nous obtenons. Nous observons du texte présent dans des images et quelques problèmes de cheminées. Les résultats sur ces mises en page complexes sont néanmoins encourageants car la plupart des paragraphes ont été segmentés et labélisés correctement.

Nous avons également participé avec la même méthode à la compétition sur la segmentation des documents présentés à ICDAR2017 [ClAP17] où nous avons obtenu avec la première méthode de segmentation pure, un taux de réussite de 81,15 %, nous classant 5 sur 7 mais avec un résultat proche du second 83,96 %, le meilleur ayant un résultat de 92,32 %. En considérant la segmentation plus la classification nous obtenons un score de 78,70 % nous laissant toujours sur la 5$^{\text{ème}}$ place du podium.



(a) (b) (c) (d)

Figure 69 – Exemples de résultats de notre méthode pour l'extraction de la mise en page obtenue sur la base PRImA (légende couleur : bleu foncé : texte, bleu clair : image et vert : séparateur).

Tableau 6 - Évaluation des résultats (comparaison des différents éléments composant la mise en page du document grâce à la mesure PRImA).

	Image (en %)	Séparateur (en %)	Texte (en %)
Dice	36,46	27,06	40,02
Fraunhofer	61,83	84,51	82,37
REGIM-ENIS	54,42	74,53	15,44
Tesseract	52,95	69,42	73,24
Felhi *et al.* [FeTS14]	67,96	78,46	89,53
Méthode proposée (2017)	95,3	94,3	81,35
Méthode proposée (2019)	95,3	94,3	84,83

3.5 Synthèse et discussions

Dans ce chapitre nous avons présenté une méthode pour extraire la mise en page des documents fondée sur la dualité entre le fond et la forme que nous décrivons grâce aux segments présents dans ceux-ci. Cette description est réalisée dans un nouvel espace, celui des traits, espace obtenu par les transformées que nous avons décrites dans le chapitre précédent. Notre méthode n'utilise pas d'apprentissage supervisé.

La première étape de l'extraction de la mise en page peut être encore améliorée. En effet, la détection de tableaux se fait à partir des traits détectés, et ne considère que les tableaux entièrement matérialisés. Or, grâce à l'analyse de la position des traits, en considérant leurs alignements et leurs espacements, il est également possible d'extraire également les tableaux semi-matérialisés. De plus, notre méthode pour agréger les cellules en tableaux peut également être améliorée en prenant en compte des caractéristiques de formes ce qui nous permettrait d'obtenir un meilleur rappel. L'étape de correction de la segmentation des régions de textes pourrait également être améliorée en affinant la recherche de l'espacement inter-lignes ou inter-mots permettant ainsi d'améliorer la détection en paragraphes.

Nous avons présenté notre méthode pour extraire la mise en page des documents et nous l'avons évalué selon un critère de qualité. Le chapitre suivant portera sur la sécurisation des documents. Nous présenterons différentes méthodes permettant de sécuriser un document et leurs critères d'évaluation.

Chapitre 4
Sécurisation des documents hybrides

Sommaire

Résumé

Dans ce chapitre, nous nous intéressons à la sécurisation des documents hybrides. Après un rapide aperçu des méthodes portant sur ce thème, nous présenterons le système de sécurisation des documents fondé sur le contenu du projet SHADES en le positionnant par rapport aux méthodes de sécurisation existantes. Nous définirons également les notions en lien avec la sécurisation, telles que l'égalité, la robustesse et la stabilité d'un algorithme. Enfin, nous évaluerons notre proposition de méthode d'extraction de la mise en page en fonction de ces derniers critères : la stabilité et la robustesse.

4.1 Introduction

La motivation du projet SHADES, était de développer un système fondé sur le contenu ayant pour objectif de sécuriser les documents. Ce projet s'inscrit dans un contexte où la dématérialisation est de plus en plus présente dans notre société pour l'échange de documents. Diverses questions juridiques se posent :

- un document numérique copié peut-il être considéré comme le même document que l'original ?
- la modification de la présentation d'un document en cours d'échange crée-t-elle une nouvelle version de ce document ?

Pour répondre à ces questions, les acteurs du projet SHADES ont introduit la notion de document « hybride », traduisant à partir d'un document son passage alternatif de document électronique à document papier par impression ou de document papier à document électronique par numérisation. Ainsi, nous considérons le contenu du document indépendamment de sa forme (papier ou numérique). Dans la vie de tous les jours, ce qui importe aux titulaires de documents ou à ceux qui y sont confrontés, c'est bien le contenu de ceux-ci et non la forme qu'ils peuvent prendre. Or, dans le contexte actuel, la validité d'un document est intimement liée à son contenu et à son contenant (numérique et papier). Se dégager de cette contrainte ouvre le champ des possibles.

Dans notre recherche, nous considérons que deux images de documents sont des instances du même document hybride si elles ont le même contenu, et sont des instances de deux documents hybrides différents si le contenu des deux documents diffère. Juridiquement, la question du document numérique est encore récente. Dans le domaine du droit, il n'existe qu'un seul et unique document matériel, tous les autres sont des copies. Deux types de copies peuvent ainsi être distinguées :

- la copie « conforme » est une copie dont tous les éléments sont présents mais pas forcément sous la même forme. Ainsi, si je recopiais à la main mon manuscrit de thèse sans rien oublier, j'aurais une copie « conforme » de celle-ci ;
- la copie « fidèle » est une copie dont la forme et le fond (au sens littéral) sont exactement les mêmes (NORME NF Z42-026)[9]. Ainsi, si je photocopie ma thèse, j'obtiendrai une copie « fidèle » de celle-ci.

[9] https://norminfo.afnor.org/norme/nf-z42-026/definition-et-specifications-des-prestations-de-numerisation-fidele-de-documents-sur-support-papier-et-controle-de-ces/116714 visité le 5 mars 2018.

Bien qu'il y ait un intérêt dans la copie conforme car elle conserve le contenu du document, les difficultés techniques entraînées et que nous considérons dans la Section 4.3 ont amené à nous limiter à la copie « fidèle ». Une modification de mise en page liée à l'utilisation de deux versions différentes d'un traitement de texte ne rentre pas dans le cadre de ce travail. Les deux images ne seront pas associées à un unique document hybride.

Avec la démocratisation des ordinateurs et des logiciels de traitement d'image, il est devenu extrêmement simple de modifier une image de document. Il existe différents moyens de modifier l'image frauduleusement, par exemple :

- CPI : copier des éléments dans un document et les coller à l'intérieur de ce même document par exemple pour modifier une valeur, une ligne de budget, *etc.* ;
- CPO : copier des éléments dans divers documents et les coller dans un autre document par exemple en introduisant ou modifiant un logo ;
- IMI : créer du texte en imitant la police de caractère que l'on veut remplacer ou ajouter par exemple pour modifier une adresse ;
- CUT : supprimer une partie de texte (et la remplacer par du fond) par exemple pour supprimer une information.

La fraude documentaire est gérée par la Délégation Nationale à la Lutte contre la Fraude (DNLF). Malheureusement, ces modifications ne sont pas les seules que les fraudeurs font subir au document. Notre objectif étant de vérifier la fidélité des versions des documents tout au long de leur cycle de vie, nous devons prendre en compte ces changements. Les versions peuvent se dégrader en fonction des conditions de conservation des documents. De plus, un document peut être imprimé par différents types d'imprimantes possédant des qualités d'encre différentes. Les différentes instances d'un document hybride ne seront pas strictement identiques les unes par rapport aux autres car des modifications naturelles apparaissent. Pour clarifier notre propos, nous avons classifié selon deux critères les modifications que peuvent subir un document au long de sa vie (cf. Tableau 7). La première classification en colonne clarifie la nature du changement et la deuxième permet de présenter nos objectifs représentés par un code couleur. Les déformations sensibles pour les utilisateurs sont en rouge, celles qui dépendent du contexte sont en violet et les déformations qui sont naturelles et par rapport auxquelles nos méthodes doivent être robustes sont indiquées en vert. À partir de cette section, nous ne considérerons comme « modification » que les modifications qui affecteraient le document d'un point de vue « légal ». Ce sont ces modifications que nous avons considérées comme sensibles (en rouge) dans le Tableau 7.

Nature	Naturel	Acquisition	Imprimante	Fraude
Modifications	- Taches (de café, d'encre, d'eau …) Papier froissé, déchiré - Partie manquante - Couleurs altérées	- Poussières - Résolutions différentes - Alignement - Bourrage imprimante	- Résolutions différentes - Alignement différent dû à l'imprimante	- Modification d'éléments - Suppression d'éléments - Ajout d'éléments - Tache d'encre

Tableau 7 - Classification des modifications pouvant survenir sur un document en fonction de leur nature (légende couleur : sensible : rouge, dépend du contexte : violet et robuste : vert).

Ce chapitre dédié à la sécurisation des documents s'organise de la manière suivante. Nous commencerons par dresser un état des lieux des techniques de sécurisation existantes pour les documents (cf. Section 4.2), puis le processus fondé sur le contenu proposé dans le projet SHADES pour sécuriser les documents hybrides sera présenté (cf. Section 4.3). Nous définirons les notions permettant d'évaluer les performances de nos algorithmes (cf. Section 4.4), puis les méthodes d'évaluation en fonction de ces notions (cf. Section 4.5). Pour finir, nous présenterons nos résultats d'extraction de la mise en page selon l'axe de la sécurité (cf. Section 4.6), et nous terminerons par une conclusion (cf. Section 4.7).

4.2 Aperçu des méthodes de sécurisation des documents

Dans cet état des lieux, nous opposons deux méthodes de sécurisation. Les méthodes de sécurisation s'attachant à protéger *a priori* les documents des modifications frauduleuses, ce sont les méthodes « actives » et celles qui vont essayer de sécuriser les documents en cherchant de potentielles modifications frauduleuses, qui sont des méthodes dites « passives ».

4.2.1 Sécurisation spécifique des documents – approche active

Dans le domaine de la sécurité plusieurs critères peuvent être considérés en fonction des besoins et de l'utilisation recherchés :

- la confidentialité : « le fait de s'assurer que l'information n'est accessible qu'à ceux dont l'accès est autorisé »[10], c'est-à-dire ceux qui détiennent une clé ;
- l'authentification : elle garantit l'identité de l'émetteur ;
- l'intégrité : elle garantit que l'information n'a pas été modifiée.

Lorsque la sécurisation vient de la constitution elle-même du document, il existe deux niveaux de sécurisation. La sécurisation du document pour que seules les personnes habilitées

[10] définie par l'Organisation internationale de normalisation (ISO).

puissent le consulter (confidentialité) et la sécurisation du document où le document doit être authentifié mais rester visible à toute personne souhaitant l'observer (authentification).

Dans un premier temps, nous présenterons les moyens de sécurisation majoritairement mis en œuvre dès la création d'un document. Ceux-ci sont généralement liés à la nature du contenant, du support initial du document. Ainsi, les techniques qui sont permises par le support papier ou plus généralement par le support matériel ne sont pas dématérialisables. Nous entendons par là que les acquérir de manière numérique leur fait perdre toute leur qualité de sécurisation, comme, par exemple, les filigranes qui ne peuvent être correctement scannés. De même, réciproquement, les techniques de sécurisation permises par les supports numériques ne sont pas toutes matérialisables, c'est-à-dire que nous ne pouvons pas les imprimer. En effet, ces moyens utilisent généralement des méta-données. La sécurisation se fait par des données cachées, invisibles dans le document. Cependant, il existe des techniques de sécurisation qui fonctionnent dans les deux univers. C'est-à-dire que la sécurisation reste intacte dans le domaine matériel et dans le domaine numérique. Nous pouvons donc alternativement scanner et imprimer le document sans perdre le moyen de sécurisation.

➢ Sécurité liée au support matériel

Les supports matériels de l'écriture existent depuis, au moins, l'invention de l'écriture et, dès cette époque, la question de l'authentification des messages s'est posée. Les moyens les plus anciens sont les scellés et les sceaux. Il existe deux types de sceaux : celui, solide qui va figer une information, garantissant l'intégrité du contenu ou du contenant avant que le sceau ne soit brisé (cf. Figure 70) et celui, réalisé à l'encre qui est déposé sur le support.

Le sceau à la cire, permet de marquer d'une empreinte une matière qui durcit avec le temps. Cette empreinte a pour but d'authentifier un document et de permettre de déterminer si le document a été altéré ou divulgué à des personnes non habilitées. Le plus vieux sceau conservé au musée du Louvre, un sceau cylindre dit du Roi-prêtre, est daté de la 2ème moitié du IVème millénaire avant JC. Ces types de sceaux étaient précédés par les cachets plats apparus au Vème millénaire[11] avant JC.

Dans le domaine physique, lorsque l'on veut garantir la confidentialité, on sécurise le contenant dans lequel est le document, que ce soit physiquement en interdisant l'accès ou en utilisant la loi *via* le biais d'un contrat promettant des dommages et intérêts si le document était diffusé.

[11] https://www.louvre.fr/oeuvre-notices/sceau-cylindre-du-roi-pretre visité le 5 mars 2018.

Figure 70 - Illustration d'une lettre cachetée par un sceau.

Il est souvent nécessaire de garantir l'authentification d'un document sans toutefois cacher le contenu de celui-ci, c'est le cas de la sécurisation des diplômes, des cartes d'identité, des billets de banque, *etc.* En effet, ce sont des documents très sensibles dans notre société. Parmi les techniques utilisées, nous pouvons noter le filigrane, une technique permettant de voir un dessin par transparence sur du papier. Les types de support ou d'encre peuvent également être des moyens de sécuriser les documents. C'est par exemple le cas des billets de banque qui utilisent un papier « ferme et craquant ».

Ces méthodes sont sensibles aux « volés vierges », c'est-à-dire des documents authentiques volés avant leur personnalisation, avant l'introduction d'un contenu. Mais elles ont l'avantage de garantir une sécurité des documents, en permettant à des néophytes d'identifier la validité d'un document. C'est particulièrement le cas du billet de banque pour lequel le ministère de l'Économie et des Finances a créé une vidéo didactique sur sa chaine YouTube pour facilement vérifier la validité d'un billet[12].

Malheureusement, ces moyens n'existent que dans le domaine « matériel ». La sécurité n'est plus valide lorsque le document est dématérialisé, du moins un tel document dématérialisé perd son authenticité.

➢ Sécurité liée au support numérique

Nous allons rapidement et sans objectif d'exhaustivité, présenter dans cette section les méthodes de sécurisation de documents dématérialisés. Dans cette catégorie nous pouvons citer toutes les méthodes de cryptographie « moderne », utilisant des services numériques pour sécuriser des messages, c'est-à-dire des documents de toutes natures (c'est-à-dire les images, les tableurs, les vidéos, *etc.*). Comme pour les documents matériels, il y a deux types de sécurisation, les méthodes qui limitent l'accès aux personnes autorisées et celles qui protègent l'intégrité des messages mais laissent les documents accessibles à tous.

[12] https://www.youtube.com/watch?v=qX7dv4V9Yq4 visité le 5 mars 2018.

Les algorithmes de la cryptographie symétrique font partie de la première catégorie. Ils permettent le chiffrement et le déchiffrement des messages grâce à une clé privée (cf. Figure 71). L'accès aux messages est réservé à ceux qui connaissent la clé. Le rapport de NIST *(The National Institute of Standards and Technology)* [Bark16] recommande les algorithmes suivants : *Triple Data Encryption Algorirhm (TDEA)* [BaMo17] et *Advanced Encryption Standard (AES)* [DaRi02].

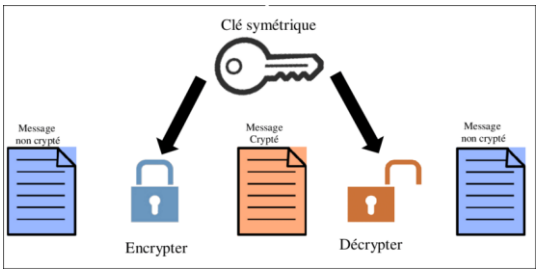


Figure 71 - Principe de fonctionnement des algorithmes de chiffrement symétrique.

Les fonctions de hachage permettent d'obtenir un code, appelé en anglais *digest*. Une fonction de hachage permet d'associer une entrée de taille arbitraire, correspondant à une suite de bits pouvant être une image de document ou un document dans n'importe quel format, à un *digest*. Elles possèdent une ou plusieurs propriétés de sécurité. Parmi les propriétés de sécurité existantes on trouve les suivantes :

- résistance aux attaques en préimage[13] : difficulté pour un attaquant connaissant un *digest* de retrouver le message ;
- résistance aux attaques en seconde préimage : difficulté pour un attaquant connaissant un message de trouver un autre message possédant le même *digest* ;
- résistance aux attaques de collisions : difficulté pour un attaquant de trouver deux messages distincts donnant le même *digest*.

En utilisant une fonction de hachage appropriée, un adversaire ne pourra pas porter atteinte à l'intégrité des données. La fonction standard est SHA256, qui donne un *digest* de taille 256 bits. Ces méthodes sont, entre autres, utilisées pour obtenir une signature électronique d'un document.

Le tatouage numérique (appelé en anglais *watermarking*) permet également de sécuriser un document numérique. Il permet d'incruster dans l'image des données permettant une sécurisation grâce à un *watermarking encodeur* créant ainsi une « *Watermarked image* ». Une

[13] Un élément dans l'ensemble de départ lié à un certain élément de l'ensemble d'arrivée est appelé préimage de cet élément.

étape de décodage permet de restaurer l'image et de localiser l'éventuelle fraude (cf. Figure 72). Il existe plusieurs types de *watermarking* :

- les systèmes dotés de capacités d'authentification exactes sont appelés *Fragile watermarking*. Nous pouvons citer dans ces méthodes, la méthode développée dans [ZhWa07], fondée sur une analyse statistique de l'image ;
- les systèmes capables de tolérer un post-traitement modéré sont appelés *Semi-fragile watermarking* [FrGo00] ;
- les systèmes capables de récupérer le contenu d'origine d'une image altérée sont appelés *Self-recovery watermaking* [HCTM12].

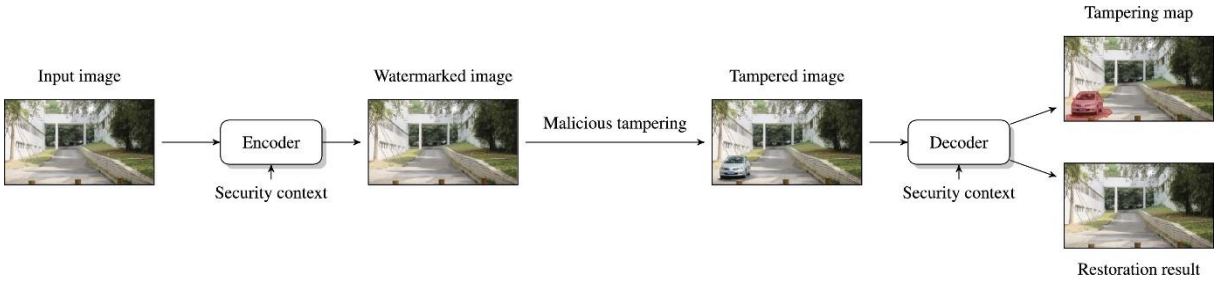


Figure 72 - Schéma du work-flow du watermarking extraite de [Koru17]. *Une version protégée d'une image est générée par un watermarking encodeur ; un décodeur correspondant vérifie le watermarking intégré et effectue une analyse pour localiser la falsification illicite ou restaure l'aspect original de l'image, en fonction des informations de référence disponibles dans le watermark.*

Le principal défaut de ces méthodes dans le cadre de notre recherche, tient dans ce qu'elles ne résistent pas à l'impression. Elles n'ont de sens que dans le domaine numérique. Ce principe de sécurisation dépend donc de la nature du support et ne peut pas basculer sur un autre. Cette approche n'est donc pas adaptée aux documents hybrides.

➢ Sécurité indépendante des supports

Après avoir montré quelques techniques de sécurisation pour des documents matériels et d'autres immatériels, nous verrons dans cette partie des méthodes de sécurisation pouvant exister dans les deux états, c'est-à-dire que l'on peut considérer que le document soit scanné ou imprimé.

Les premiers moyens de sécurisation qui résistent aux scans sont le sceau à encre et la signature manuscrite. Le sceau considéré ici, est, contrairement au sceau (utilisant de la cire) celui présenté dans la section matérielle, un tampon qui fournit un motif pouvant être numérisé. Celui-ci sert également à appuyer l'authenticité du document et est apparu très tôt dans l'histoire. Le sceau est encore utilisé par les administrations publiques. La signature manuscrite

est plus tardive. Elle témoigne d'une certaine généralisation de l'écriture, mais il est admis qu'elle caractérise de manière unique un individu. Elle permet, à moindre frais, d'authentifier l'auteur d'une lettre ou l'approbation d'un individu sur son contenu. Ces moyens ne sont pourtant pas considérés comme des moyens de sécurisation fiable car même s'ils résistent à la dématérialisation, ils perdent en sécurité. En effet, de nombreuses caractéristiques chimiques et physiques qui permettent de les authentifier avec plus de certitude ne se retrouvent pas dans la version numérique. Ils sont plus difficiles à authentifier par des experts humains. De plus, le fait que la signature ou le sceau soit authentique ne garantit pas que le reste du contenu n'ait pas été modifié *a posteriori*.

Dans le cadre de la sécurisation des documents, une solution technologique qui résiste aux scans et à l'impression est le 2D-Doc mise en place en 2013. Il s'agit d'une solution que met en avant le gouvernement français[14] pour protéger les documents. Cette solution a l'avantage de pouvoir sécuriser des champs dans un document mais son inconvénient réside dans le fait que les documents doivent être nativement numériques. Le 2D-Doc sécurise seulement certaines informations dans un code barre à 2 dimensions visibles dans le document (cf. Figure 73). Pour vérifier si le document n'a pas été modifié, il convient alors de décoder ce code barre. Les données de définition du code 2D-Doc sont utilisées pour extraire les informations du code (Message, Signature, Annexe éventuelle). Puis ces données sont comparées avec celles mentionnées dans la partie visible du document sur lequel il a été apposé.

Cette méthode possède deux inconvénients principaux. D'une part il faut que le document ait été créé ainsi. Par ailleurs, elle ne sécurise que certains champs et non le document dans sa globalité.

Figure 73 - Exemple de 2D-DOC.

➢ Conclusion

Le défaut majeur de ces méthodes de sécurisation, hormis les méthodes de cryptographie, est qu'elles sont natives, c'est-à-dire que le document doit avoir été créé en prenant en compte ces questions de sécurité. Cela signifie que ces méthodes de sécurisation

[14] https://ants.gouv.fr/Les-solutions/2D-Doc visité le 8 mars 2018.

sont réservées à un petit groupe de documents conçus comme tels et donc ne permet pas de sécuriser tous les documents.

4.2.2 Sécurisation des documents par la recherche ou la détection de modifications frauduleuses – approche passive

Dans cette partie, nous nous intéresserons à la deuxième catégorie de méthodes développées pour lutter contre la fraude : ces méthodes recherchent la présence de la fraude dans les documents *a posteriori*. Nous commencerons par discuter des méthodes utilisées pour les documents uniquement graphiques, puis des méthodes se fondant sur une analyse au niveau de l'information textuelle.

- **Analyse des éléments graphiques d'un document**

Dans cette partie, nous considérerons plutôt les documents constitués d'images naturelles qui n'ont généralement jamais été imprimées sur support papier. La problématique liée au *print & scan* (impression et numérisation) n'est pas considérée. La modification frauduleuse, ici, est très fréquente. De nombreux utilisateurs réalisent des photomontages destinés en particulier au divertissement. Les modifications dans ces cas sont généralement assez grossières (il n'y a pas de doute sur le fait de visualiser un faux). Cette pratique est plus problématique quand celui qui modifie l'image, le fait dans un but malveillant ou de tromperie. Par exemple, pour calomnier une personne ou, dans les pires cas, pour falsifier des preuves. Cette problématique a, depuis longtemps, été prise en compte. Mais l'amélioration des techniques et logiciels de traitement d'images fait qu'il est toujours plus facile de modifier de manière très naturelle et précise les images, il est alors aussi plus difficile de détecter la fraude. Nous pouvons citer dans les avancées technologiques, les permutations intelligentes de visage en temps réel dites *Face2Face* [TZCM16] qui permettent de simuler les mouvements faciaux d'une personne.

Ce domaine est très prisé. En anglais, on parle de *document forensics*. La recherche de faux fait partie des *passive forensics analysis*. Nous pouvons citer les nombreux état de l'art sur ce sujet tel que [Koru17], [CoPV15], [HLHN15], [Piva13], *etc.*

- **Analyse de la partie textuelle des documents**

Ces méthodes sont fondées sur l'analyse des éléments ayant trait au texte, que ce soit en analysant le comportement de la texture constituée par la présence de texte et celle du fond autour des caractères, ou en analysant le contenu textuel lui-même.

Les documents imprimés possèdent une texture unique sur chacun des caractères. Trouver deux fois la même texture sur un même caractère imprimé en deux endroits différents est un indice d'une fraude par copier-coller de type CPI. Cette technique est utilisée par [BGTF13], [AbBö16].

La détection du type d'impression est également un moyen de sécurisation. Par exemple la méthode de [LaMB06] cherche à détecter les documents possédant plusieurs types d'impression. Ainsi, grâce à un algorithme d'apprentissage (SVM), les caractères sont classés en différentes catégories suivant leur type d'impression (encre ou laser). La mise en évidence de l'absence d'homogénéité des caractères d'impression dans un même document souligne une altération.

Les méthodes sémantiques se basent sur la cohérence du texte. C'est l'objet de la thèse de Chloé ARTAUD [Arta19]. En liant les informations par une ontologie, nous pouvons vérifier leur vraisemblance. Par exemple, sur une base de tickets de caisse, on peut vérifier si la somme des coûts des articles est égale au prix total, si une quantité vendue multipliée par le prix unitaire est égale au prix noté sur le ticket, si la date et l'heure existent, *etc.*

➢ **Conclusion**

Une compétition a été organisée en 2018 dont la tâche était de détecter la fraude documentaire [ASDO18]. L'une des difficultés majeures de ce domaine est la création d'un corpus réaliste. En effet, les entreprises ne souhaitent pas donner les documents véritablement fraudés. Le corpus de la compétition contenait principalement des tickets venant de différents magasins et restaurants et contenaient des images fraudées par des doctorants et post-doctorants de l'équipe des organisateurs selon les principales méthodes de modifications (CPI, CPO, IMI, CUT, *etc.*).

L'étude a également exploré les résultats de la détection de fraude humaine en demandant à 5 humains de détecter la fraude. Leurs résultats étaient inférieurs à ceux des méthodes informatiques avec une précision à 0,75 et un rappel à 0,5 pour le meilleur. Cette expérience bien que menée avec un échantillon réduit d'êtres humains, confirme que les êtres humains sont plutôt mauvais dans une tâche de détection d'éléments fraudés, cela renforce la nécessité de concevoir des outils pour détecter la fraude.

Le principe du processus de sécurisation développé dans le cadre du projet SHADES est situé dans la catégorie des méthodes actives en combinant plusieurs avantages : il n'est ni

lié à un support, ni natif (même s'il fige une version dans le temps que l'on considérera comme un original qui servira de référence pour des comparaisons).

4.3 Sécuriser les documents grâce au processus de SHADES

Le processus de sécurisation fondé sur le contenu proposé dans le cadre du projet SHADES est fondé, comme son nom l'indique, sur un hachage sémantique.

Pourquoi parle-t-on de « sémantique » ? La sémantique est la branche de la linguistique qui étudie le sens, la signification. Il s'agit dans notre cas d'un abus de langage, les techniques automatiques actuelles ne permettent pas de valider si deux textes ont le même sens mais c'est bien par ce principe que l'idée sous-jacente du projet est née. Peu importe la qualité, la résolution ou la luminance de l'image d'un document, seul le contenu est pris en compte. Les difficultés sont donc de trouver pour chaque instance d'un document hybride, le même contenu et d'extraire ce contenu. En effet, si pour le texte nous pouvons de ce dernier penser que le contenu correspond aux lettres qui le composent, ce n'est pas le cas des autres médias comme les images ou les logos, *etc.*

Le deuxième terme de « hachage » est directement lié à l'aspect sécurité. Il est expliqué par la problématique des entreprises. Comme nous l'avons évoqué dans la Section 4.2.1, les méthodes de hachage permettent d'obtenir un code irréversible lié aux données. Si les données sont modifiées et ce, même si la modification est infime, alors le code obtenu sera complétement différent de celui obtenu avant la modification. Nous pouvons ainsi sauvegarder une trace d'un changement sans compromettre la confidentialité au niveau du contenu des documents. Cette contrainte est importante pour les entreprises, elle protège les clients dans le cas d'une attaque où les hackeurs récupèreraient leurs bases de données. Cela se place ainsi dans un contexte plus général de sécurisation des données. Selon la loi, c'est à l'entreprise de garantir à l'utilisateur la protection des données.

Le processus de sécurisation proposé dans le cadre du projet SHADES permet donc d'obtenir pour chaque document hybride un code lié à son contenu. Code que nous noterons par le terme « cachet », à la place du terme « signature » présent dans le nom du projet. En effet, SHADES signifie « Hachage sémantique pour la signature électronique avancée de document ». Il s'agit d'un abus de langage qui est préjudiciable aussi bien du point de vue informatique que du point de vue juridique. En effet le terme de « signature » sous-entend que nous authentifions le contenu et l'auteur du document. Comme nous l'avons expliqué plus tôt, le projet SHADES n'a pas vocation à authentifier le document ou l'auteur. C'est pourquoi nous avons donc préféré le terme de « cachet ».

Cela étant posé, nous pouvons présenter le processus de sécurisation. Ainsi pour chaque instance de document, une image de document, les régions d'intérêt sont extraites et chaque région est labélisée selon le type de média auquel elle appartient (texte, graphique, *etc.*). La méthodologie d'extraction de la mise en page présentée dans le chapitre 3 a ainsi été proposée dans ce cadre. Ces régions sont ensuite décrites en fonction de leur label, car les informations identifiant le contenu dépendent du média considéré. À partir des informations extraites, il faut prendre en compte des caractéristiques invariantes aux modifications naturelles. Ces descripteurs résultants sont liés aux labels qu'ils décrivent. Par exemple, dans le texte, ce sont les lettres qui le composent, en complément nous pouvons ajouter la police de la fonte et la taille. Ces informations permettent de reproduire le texte sans avoir besoin de sauvegarder l'image. Une image, contrairement au texte, dépend de l'instance d'un document hybride. Il faut ainsi trouver une autre façon d'en extraire des informations. Nos partenaires du L3i ont proposé une méthode nommée « ASYCHA » permettant d'extraire les informations de l'image (et des logos) ne dépendant pas de l'instance du document que l'on étudie. Cette méthode est un algorithme de hachage perceptuel calculé sur une représentation grossière de l'image. Les descriptions de chaque région auxquelles

on ajoute la description structurelle de la mise en page (cf. Section 4.5.3) sont alors hachées selon un algorithme défini et nous obtenons un code qui sera (idéalement) le même pour chaque instance de document mais différent si le document a été modifié. La Figure 74 présente la chaîne de traitements du processus de sécurisation issu du projet SHADES sur trois instances d'un document hybride dont l'un a été modifié. Les versions 1 et 2 obtiennent à la fin du processus le même cachet, alors que le document qui a subi une modification obtient un cachet différent.

Le processus de sécurisation initial du projet SHADES a été revu suite aux limitations techniques dues à la stabilité du résultat obtenu lors de l'extraction des régions d'intérêt. Il est extrêmement difficile d'obtenir un résultat strictement identique (tels que nous le définissons dans la Section 4.5) pour toutes les instances d'un document, une fonction de hachage globale n'est donc pas adaptée puisque les algorithmes fournissent des régions différentes en fonction de l'instance du document. Néanmoins les propositions restent acceptables pour une bonne interprétation du document dans les usages traditionnels. Un hachage plus local, qui ne hacherait pas d'un coup tous les descripteurs mais description par description, permettant de cibler la modification semble plus adapté pour détecter un changement frauduleux. De plus, dans l'état actuel, nous ne pouvons renoncer aux originaux des documents comme c'était initialement prévu et ne comparer que les codes obtenus car la fiabilité du résultat impose une

vérification manuelle de celui-ci. Ainsi l'outil proposé par le projet SHADES est pour l'instant un prototype pour créer une alerte qu'un humain devra vérifier manuellement.

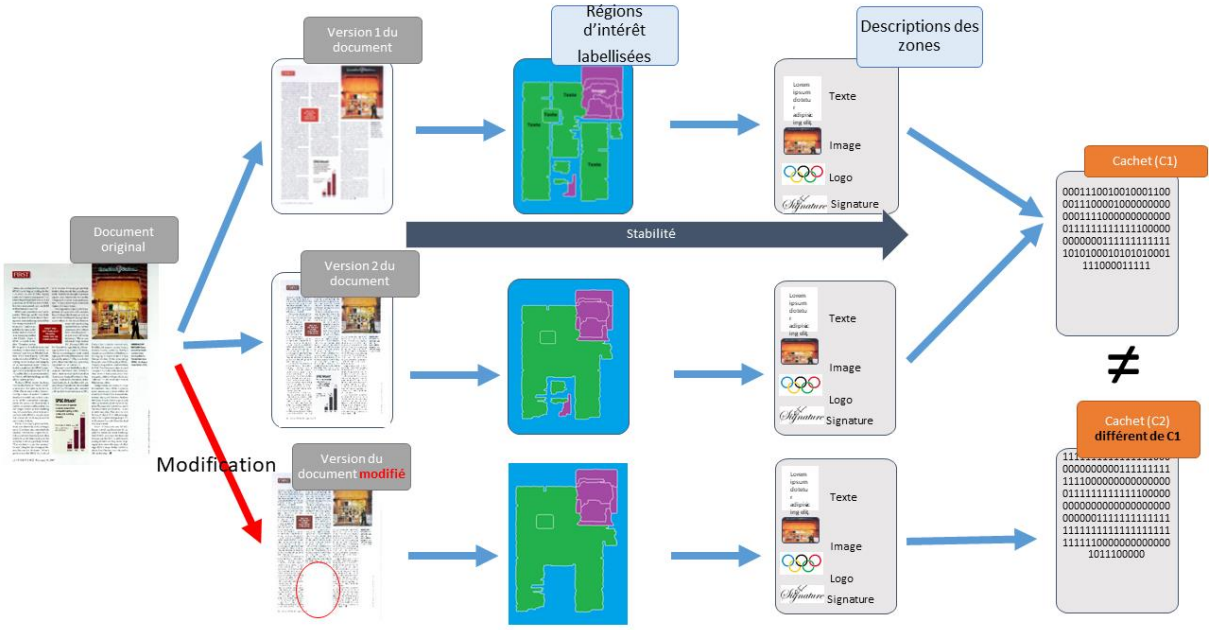


Figure 74 - Chaîne de traitements du processus développé dans le cadre de sécurisation du projet SHADES sur trois instances d'un document hybride.

Nous avons évoqué dans cette section que deux documents pouvaient avoir le même contenu, nous avons également parlé de stabilité d'un algorithme mais ces notions peuvent être prises sous différents points de vue et il convient de les clarifier, ce que nous nous proposons de faire dans la section suivante.

4.4 Définitions préliminaires

Tout d'abord nous considérons la notion de stabilité car, comme nous l'avons évoqué dans l'introduction, le but du projet SHADES est de fournir un code unique pour chaque document hybride, indépendant de l'instance image considérée. Néanmoins, le passage du fichier électronique à une version papier ou l'inverse s'accompagne bien souvent de dégradations. Ces altérations ne doivent pas modifier le caractère légal que peut avoir le document sinon une des instances de celui-ci ne sera pas considérée comme appartenant au même document hybride. Par contre, il est nécessaire de pouvoir identifier des changements intentionnels qui pourraient avoir eu lieu entre la création du document et le moment où il est observé au cours de son utilisation, et qui modifieraient le document selon le point de vue « légal ». C'est le respect de ces contraintes que nous désignons par le terme de stabilité de la méthode. La stabilité est donc la capacité à analyser un document même dans des conditions

non optimales. La notion de stabilité s'applique à toute action ou méthode. Une action/méthode n'est stable que par rapport à un changement de conditions (bruits, dégradations naturelles, *etc.*).

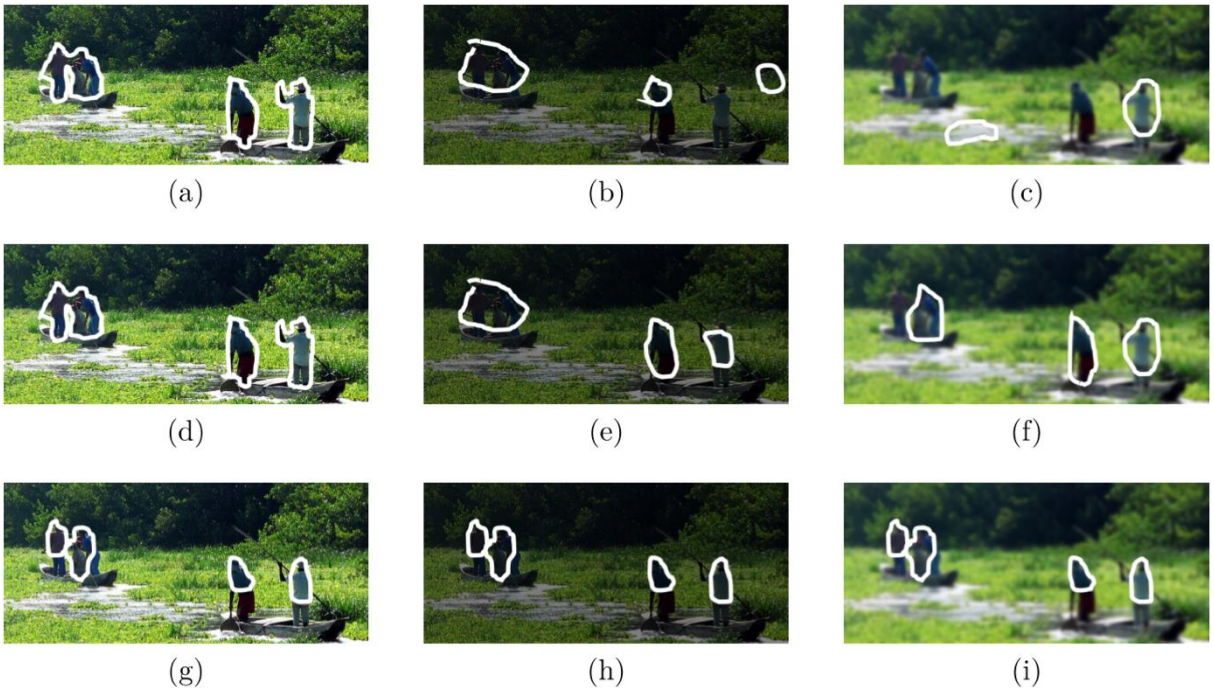


Figure 75 - Illustration des notions considérées pour la stabilité tirée de [Eske17]. *Comparaison des résultats de plusieurs algorithmes (1ère ligne : un algorithme considéré comme précis, 2ème ligne : un algorithme considéré comme robuste, 3ème ligne : un algorithme considéré comme stable) sur trois instances d'une image (1er colonne : image normale, 2ème colonne : image assombrie, 3ème colonne : image floutée).*

La stabilité a été étudiée et définie dans la thèse de Sébastien Eskenazi [Eske17] pour un algorithme général. Sur la Figure 75, les définitions de précision, robustesse et stabilité de S. Eskenazi sont illustrées. La méthode considérée est un processus de détection de personnes dans une image, et toutes les expériences concernent la même scène avec des acquisitions différentes. Les colonnes de la Figure 75 adressent des images de qualité variable. Dans la colonne de gauche, l'image est de bonne qualité, dans la colonne centrale, l'image est sombre et dans la colonne de droite, l'image est floue. Chaque ligne de la Figure 75 considère des méthodes différentes pour extraire des personnes. Les images de la première ligne montrent que dans le cas d'une image de bonne qualité, les résultats aboutissent à la détection des quatre personnes correctement segmentées. Mais si l'image est sombre ou si l'image est floue, les résultats sont alors mauvais. L'algorithme ne détecte plus toutes les personnes (vrais négatifs), de plus il détecte également du fond comme des personnes. La deuxième ligne illustre le cas d'un algorithme **robuste**, les détections sont à peu près les mêmes, quelle que soit la dégradation provoquée par la phase d'acquisition. Sur la troisième ligne, l'algorithme détecte mal les personnes, la partie inférieure de trois d'entre eux n'est pas détectée. En revanche, dans les

trois images, les positions trouvées sont strictement aux mêmes endroits, la méthode est alors qualifiée de **stable**.

Nous avons adopté un vocabulaire légèrement différent de celui de S. Eskenazi. Il [Eske17] utilise le terme de « *Accuracy* » pour parler de qualité. Cette définition d'*Accuracy* requiert une vérité terrain pour évaluer à quel point un résultat est proche de celui escompté.

Nous abordons dans la suite les autres notions que sont l'égalité, la robustesse et la stabilité.

4.4.1 Égalité

La définition d'« égalité » selon le Larousse[15] est « Qui a la même valeur, la même durée, la même importance, *etc.*, que quelque chose d'autre ». La notion d'égalité ne peut être considérée que si les éléments sont comparables. Ainsi, nous ne pouvons comparer que des choses / objets semblables. Or, si nous considérons que deux instances de document sont égales, c'est-à-dire qu'elles ont le même contenu, elles ne sont pas pour autant strictement numériquement semblables. En effet, nous possédons pour chaque instance de document hybride une image scannée, c'est-à-dire une matrice de pixels pouvant être représentée par trois canaux si l'image est en couleur. Dans ce cas, n'importe quelle altération (volontaire ou non) entraîne une modification dans le contenu de la matrice, les matrices à comparer sont donc différentes. Les différents objets sont donc difficiles à comparer si ce n'est à dire qu'ils sont systématiquement différents.

Pour comparer deux images nous faisons alors appel à la notion de robustesse, transformant la comparaison binaire d'égalité entre deux éléments en une notion qui s'applique à une transformation et qui peut se mesurer de manière continue.

4.4.2 Robustesse

Nous ne reprenons pas ici la définition proposée par S. Eskenazi [Eske17] qui considère que la robustesse n'est pas un indicateur de performance en soi, mais met en évidence la capacité d'un algorithme à maintenir une bonne précision même avec des conditions dégradées. Par conséquent, la robustesse est liée à la capacité d'un algorithme à produire des résultats significatifs lorsque l'entrée est bruitée.

[15] https://www.larousse.fr/dictionnaires/francais/%C3%A9gal/27988?q=%C3%A9gale#27846 visité le 10 mars 2019.

Dans notre cas, pour définir la robustesse, nous reprenons la définition de « robuste ». Ce mot est un adjectif qui s'applique par définition à quelque-chose qui doit résister à des agressions et à des altérations. Nous définissons donc la robustesse comme la capacité à résister à un phénomène particulier. Nous avons évoqué dans l'introduction les modifications auxquelles nos traitements doivent être robustes (cf. Tableau 7). Cette notion de robustesse permet d'introduire et de définir la notion de stabilité.

4.4.3 Stabilité et sensibilité

La stabilité est ici la capacité d'un algorithme à fournir des résultats similaires. Ainsi dans le cadre du projet SHADES un algorithme est stable s'il est robuste à l'altération des couleurs, aux changements de résolution, *etc.* (cf. Tableau 7). La définition de S. Eskenazi [Eske17], considère que la stabilité n'exige aucune vérité terrain, c'est cette définition que nous adoptons. La stabilité requiert au moins deux résultats associés à des entrées similaires (c'est-à-dire qu'elles comportent le même nombre d'éléments, que les contenus des éléments soient identiques et que ceux-ci aient la même disposition les uns par rapport aux autres) et la mesure de la proximité de ces résultats en fonction de la proximité des entrées. La proximité permet d'évaluer la ressemblance de ces deux éléments.

La stabilité, suivant le point de vue que nous venons d'introduire peut-être validée ou non, mais elle peut aussi être mesurée par des valeurs sur une échelle continue. Si la notion est considérée de manière binaire, alors une légère différence entre les deux entrées conduit à la conclusion que l'algorithme résultat n'est pas stable. A l'heure actuelle, les algorithmes ne sont pas assez stables pour obtenir des résultats aussi bons, notamment ceux issus de l'étape de segmentation. Nous choisissons alors d'évaluer la stabilité de manière continue sur l'intervalle [0, 1].

Pour revenir à notre application, si, dans l'extraction des différents éléments constituant la mise en page de toutes les instances de documents, la conclusion est qu'il n'y a aucun élément, alors la stabilité sera excellente car nous aurons toujours le même résultat. Pour autant le système ne sera pas pertinent et surtout pas sensible. Nous avons ici introduit la notion de sensibilité. Il s'agit de considérer deux documents comme différents si ceux-ci sont effectivement différents.

En évaluant la qualité des résultats des méthodes, comme nous l'avons fait dans le chapitre précédent, nous écartons de notre considération le cas extrême où nous n'avons trouvé aucune région d'intérêt. Dans la section suivante nous présenterons des méthodes permettant d'évaluer la stabilité dans l'extraction de la mise en page des documents hybrides.

4.5 Méthodes d'évaluation de la stabilité

Il existe plusieurs mesures d'évaluation de la stabilité en fonction de différents degrés de robustesse que nous voulons considérer. Nous présenterons trois méthodes d'évaluation : les méthodes sensibles aux modifications physiques, recherchant donc une stricte égalité entre différentes entrées, les méthodes partiellement sensibles à ces modifications et une méthode insensible à ces modifications.

4.5.1 Évaluation sensible aux modifications physiques

Dans cette section, les mesures décrites ne considèrent que la stabilité spatiale, c'est-à-dire que l'on considère une superposition alignée en haut à droite entre les différentes entrées. Ces mesures sont pertinentes uniquement si les documents à comparer sont préalablement recalés entre eux avec un algorithme de recalage.

Nous considérons la mesure *(CL)* « *covering level* » qui permet de calculer le taux de recouvrement d'un élément e_1^i contenu dans un document doc_1 par rapport à un autre élément e_2^j contenu dans un deuxième document doc_2, il s'agit de l'indice Jaccard :

$$CL\left(e_1^i, e_2^j\right) = \frac{A\left(e_1^i \cap e_2^j\right)}{A\left(e_1^i \cup e_2^j\right)} \tag{29}$$

où $A(x)$ désigne l'aire d'un élément x.

À partir de cette mesure, nous pouvons définir, dans un document hybride dont on dispose de deux instances, la stabilité ST d'un élément e_1^i dans un document doc_1 par rapport aux éléments $\left\{e_2^j\right\}_{j=1}^{N_2}$ contenus dans un deuxième document doc_2 comme :

$$ST\left(e_1^i, doc_2\right) = \begin{cases} 0 \text{ si } N_2 = 0 \\ \max_{j\in[1,N_2]}\left\{CL\left(e_1^i, e_2^j\right)\right\} \text{ sinon} \end{cases} \tag{30}$$

La stabilité $Stab$ entre deux documents doc_1 et doc_2 possédant respectivement les éléments $\left\{e_1^i\right\}_{i=1}^{N_1}$ et $\left\{e_2^j\right\}_{j=1}^{N_2}$ est alors définie par :

$$Stab(doc_1, doc_2) = \begin{cases} 1 \text{ si } N_1 = N_2 = 0 \\ 0 \text{ si } N_1 \cdot N_2 = 0 \text{ et } N_1 + N_2 \neq 0 \\ \frac{1}{2}\left(\frac{1}{N_1}\sum_{i=1}^{N_1} ST\left(e_1^i, doc_2\right) + \frac{1}{N_2}\sum_{i=1}^{N_2} ST\left(e_2^i, doc_1\right)\right) \text{sinon} \end{cases} \tag{31}$$

Le score de stabilité globale SC sur un corpus d'instances d'un unique document hybride possédant un ensemble $D = \{doc_i\}_{i=1}^{N_H}$ d'images de document est alors défini par :

$$SC(D) = \frac{2}{{N_H}^2 - N_H} \sum_{i=1}^{N_H} \sum_{j=1}^{i-1} Stab(doc_i, doc_j) \tag{32}$$

Comme nous l'avons évoqué au début de ce paragraphe, ce score est sensible aux modifications physiques du document car les positions spatiales des résultats obtenus, les éléments de mise en page, sont comparées. Nous verrons par la suite une modification permettant d'intégrer une modification physique du document sans pénaliser le score de stabilité globale.

4.5.2 Évaluation partiellement sensible aux modifications physiques

Dans les mesures précédentes, les différences entre éléments sont prises en compte et se cumulent en fonction du contenu du document. De manière à pouvoir qualifier de robuste ou non l'extraction de la mise en page d'un document, il est possible d'introduire la notion de risque dans la bonne détection d'un élément. La mesure d'évaluation précédente devient binaire. Si la mesure SC (cf. équation 29) qui quantifie la différence de positionnement entre deux éléments e_1^i et e_2^j , est supérieure à un certain seuil s, nous considérons que le résultat prend la valeur 1 sinon il vaut 0, soit :

$$CL_s(e_1^i, e_2^j) = \begin{cases} 1 \text{ si } \dfrac{A(e_1^i \cap e_2^j)}{A(e_1^i \cup e_2^j)} \geq s \\ 0 \text{ sinon} \end{cases} \tag{33}$$

En utilisant cette modification dans l'équation (30), nous obtenons $ST_s(e_1^i, doc_2)$, puis en remplaçant ST par ST_s dans l'équation (31) nous obtenons $Stab_s(doc_1, doc_2)$. Pour finir par remplacer $Stab(doc_1, doc_2)$ par $Stab_s(doc_1, doc_2)$ dans l'équation (32) pour obtenir le score de stabilité $SC_s(D)$.

Cette métrique permet d'ignorer une modification du document dans le plan de l'image si celle-ci est inférieure à un certain seuil que peut choisir l'utilisateur. Mais elle possède deux inconvénients. Le seuil s ne dépend pas de la taille des éléments. Ainsi, une modification sera plus absorbée si les documents sont grands par rapport à de petits éléments. Deuxièmement, s'il n'y a pas de modification spatiale entre les documents à comparer mais que le résultat de l'algorithme contient une erreur si celle-ci est inférieure au seuil, l'erreur ne sera pas détectée. Dans la Figure 76, nous pouvons observer que si la translation a créé le même décalage dans les deux éléments celui de gauche est relativement moins impacté que celui de droite.

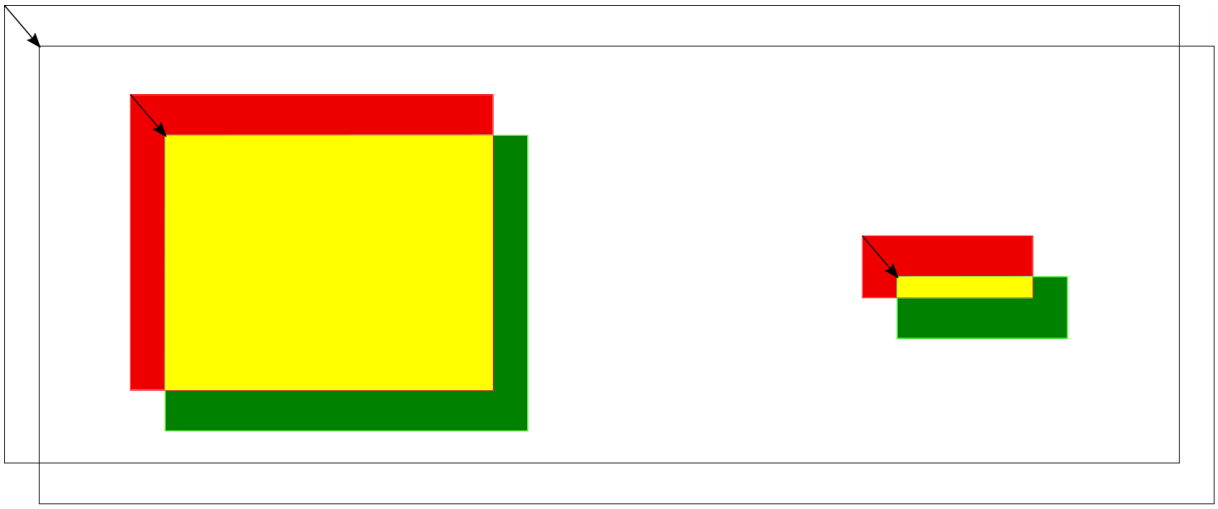


Figure 76 - Illustration de la superposition de deux instances d'un même document, ayant subi une translation, contenant deux éléments (en rouge ceux du premier élément, en vert ceux du second et en jaune la superposition des deux)

4.5.3 Évaluation insensible aux modifications physiques

Une méthode proposée par S. Eskanazy *et al.* [EsGO15] permet de comparer les éléments issus de différentes images entre eux sans utiliser de superposition. Dans cette section, chaque élément considéré est une région. Le résultat permet de savoir si l'agencement spatial est le même et dans notre cas si la mise en page est la même dans deux documents différents.

Cette mesure repose sur la comparaison entre les centres de gravité des éléments, ce qui permet de ne pas considérer le problème de la surface et des modifications géométriques des éléments. Grâce à ces points, la triangulation de Delaunay associée aux centres de gravité peut être déterminée, permettant l'obtention d'un graphe. Un tri sur les nœuds du graphe obtenu est effectué pour obtenir un ordonnancement déterministe des régions. Une comparaison stricte entre les matrices d'adjacences (tout en vérifiant le type des régions) issues du résultat de la segmentation sur différentes images de documents permet de savoir si les versions sont différentes ou issues du même document. Cette méthode est également utilisée dans le processus de SHADES pour décrire la mise en page comme l'illustre la Figure 77.

Une amélioration de cette technique a été réalisée [GPKB18] pour prendre en compte une comparaison non binaire, rendant possible la comparaison entre deux documents sur lesquels des nombres différents de régions ont été extraits. Le nouveau processus de comparaison est fondé sur une mise en correspondance locale. Au lieu d'un unique graphe, la comparaison est faite sur différents sous-graphes (cf. Figure 78). Le but est de faire correspondre les matrices d'adjacence locale entre elles (cf. Figure 79). Le pourcentage de matrices d'adjacence locale qui se correspondent est alors seuillé. Afin de préserver les relations spatiales,

la distance euclidienne entre les points d'ancrage des sous-graphes doit correspondre. Si la distance est inférieure à un seuil, les matrices d'adjacence locale seront mises en correspondance. Le seuil recommandé correspond à 10 % de la largeur du document.

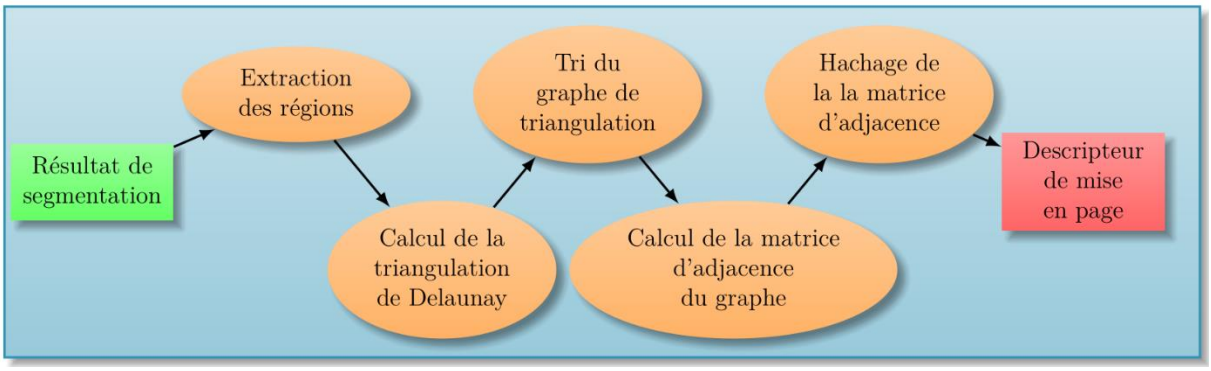


Figure 77 - Processus de calcul du descripteur de Delaunay, figure extraite de [Eske17].

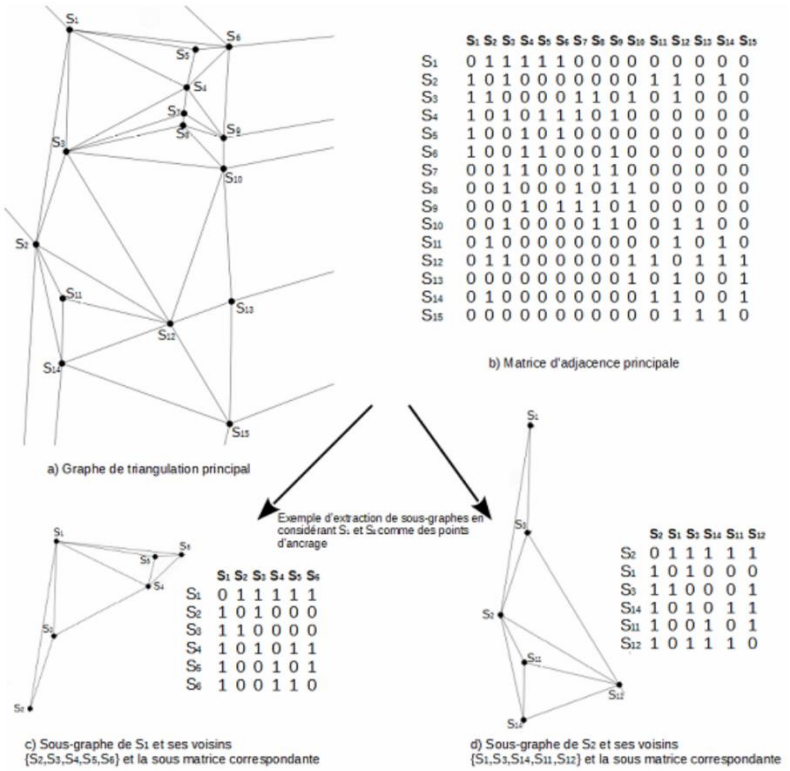


Figure 78 - Principe d'extraction des sous-graphes et calcul des matrices d'adjacence, figure extraite de [GPKB18].

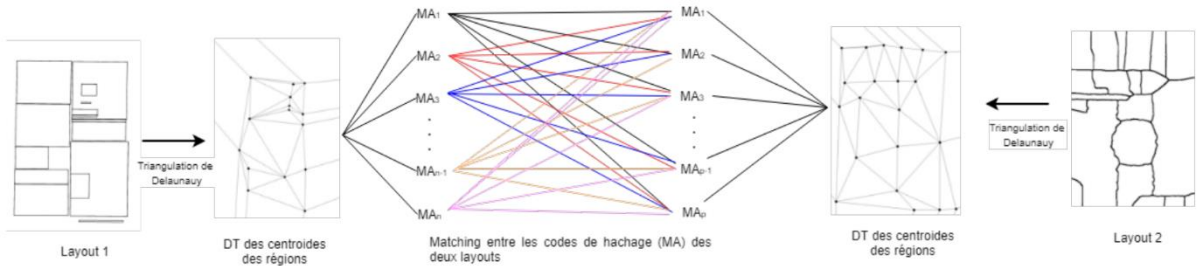


Figure 79 - Principe de la mise en correspondance entre deux mises en page, figure extraite de [GPKB18].

Dans cette section, nous avons présenté des méthodes permettant d'évaluer quantitativement la stabilité de la mise en page. Ces méthodes ont un intérêt différent permettant dans le premier cas une évaluation fondée sur des éléments plus locaux et dans le second cas une évaluation fondée sur des éléments plus globaux. En effet, la méthode choisie dans le projet SHADES, le DLD, permet d'évaluer la stabilité en termes de relation spatiale sans s'intéresser à la forme des régions. Cela a des avantages en termes de comparaison, elle peut prendre, par exemple, en compte une bordure plus ou moins large entre les différentes régions sans considérer numériquement la largeur, mais l'approche ne permet pas de considérer le problème de l'extraction d'une colonne en moins dans un tableau. Ce n'est pas un inconvénient dans le cadre du projet SHADES car la différence devra être détectée dans l'étape permettant de décrire les régions d'intérêt. Néanmoins, sans cette étape, cela offre une idée globale de la stabilité. L'autre inconvénient de cette approche est le nombre de régions traitées. En effet, si le nombre de régions est faible (égal à deux par exemple) la méthode ne permet pas de différentier deux mises en pages différentes.

Contrairement au DLD, les estimations fondées sur le *covering level*, évaluent quant à elles spatialement la stabilité. Ces estimations vont être sensibles au positionnement de la page dans l'image. Aussi, il est recommandé de réaliser un recalage des couples d'images comparées avant d'évaluer la stabilité. Le recalage permet de comparer spatialement des zones en correspondance sans ambiguïté. Cela est d'autant plus important que les zones sont de faible taille comme illustré sur la Figure 76. Une telle stratégie permet également de considérer n'importe quel nombre d'éléments.

4.6 Évaluation de la stabilité

Après avoir présenté l'intérêt de la notion de stabilité du processus d'extraction de la mise en page d'un document pour réaliser la sécurisation d'un document hybride, nous avons choisi d'évaluer les résultats obtenus avec la méthode présentée sous ce prisme de la stabilité. Nous présenterons dans un premier temps les problèmes de stabilité de l'extraction des traits et les résultats de stabilité sur les tableaux, dans un second temps, une étude liée à la stabilité de

notre segmentation et dans un dernier temps, une étude de la stabilité de la méthode de l'extraction de la mise en page.

4.6.1 Étude de la stabilité de l'extraction de tableaux

Pour évaluer la stabilité de l'extraction de tableaux, nous commençons par étudier la stabilité de la méthode d'extraction des séparateurs matérialisés (cf. Section 3.2). La stabilité est ici principalement liée à la qualité de l'acquisition et de la binarisation réalisée pour chaque instance. Les séparateurs matérialisés, en général, sont parmi les éléments les plus instables dans les documents hybrides. Cela s'explique par leur finesse par rapport à la résolution choisie. En effet, ils peuvent être représentés par un trait d'un seul pixel d'épaisseur. Ils peuvent à la fois disparaître lors de l'impression ou lors du scan. L'usure du document leur est également très préjudiciable. Ainsi, les traits peuvent, sur l'image originale, être difficilement lisibles, même pour l'œil humain. Par exemple, nous pouvons observer, sur la Figure 80, quatre instances d'un document hybride représentant une facture. Si, dans la première ligne, les images originales des instances du document nous permettent de voir tous les traits du document, nous pouvons observer, sur la deuxième ligne, le résultat de la binarisation des images, que certains traits ont disparu. Les différents zooms de l'image originale permettent de comprendre que les niveaux de gris diminuent fortement sur certaines zones et, qu'à certains endroits, de petites parties sont manquantes.

Cela est une justification supplémentaire qui nous a conduit à traiter les séparateurs matérialisés en premier dans le traitement de la mise en page, leur suppression permet d'assurer une meilleure stabilité globale par rapport aux documents hybrides.

Nous présentons ici une évaluation de l'extraction des tableaux dans le cadre de documents hybrides. Pour cela, nous utilisons le jeu de données détaillé dans le chapitre précédent (cf. Section 3.4.2), « SETSTABLE ». Avec les méthodes définies dans la Section 4.5, nous avons évalué la détection de tableaux sous l'angle de la stabilité en calculant le score de stabilité globale (cf. équation (31)) et selon le score de stabilité binaire avec un seuil pour lequel nous avons fixé deux valeurs : 0,70 et 0,80.

La Figure 81 illustre deux exemples qualitatifs de superposition de résultats réalisés sur des instances de documents hybrides. Dans le premier document hybride (cf. Figure 81 (a)), nous observons que si globalement les tableaux ont été détectés et localisés, un tableau (en haut) a été détecté à tort dans une instance ce qui diminue le score de stabilité. Dans le deuxième document hybride (cf. Figure 81 (b)), un autre tableau a également été détecté à

tort sur peu d'instances mais sur ces deux documents nous observons que les résultats sont globalement stables.

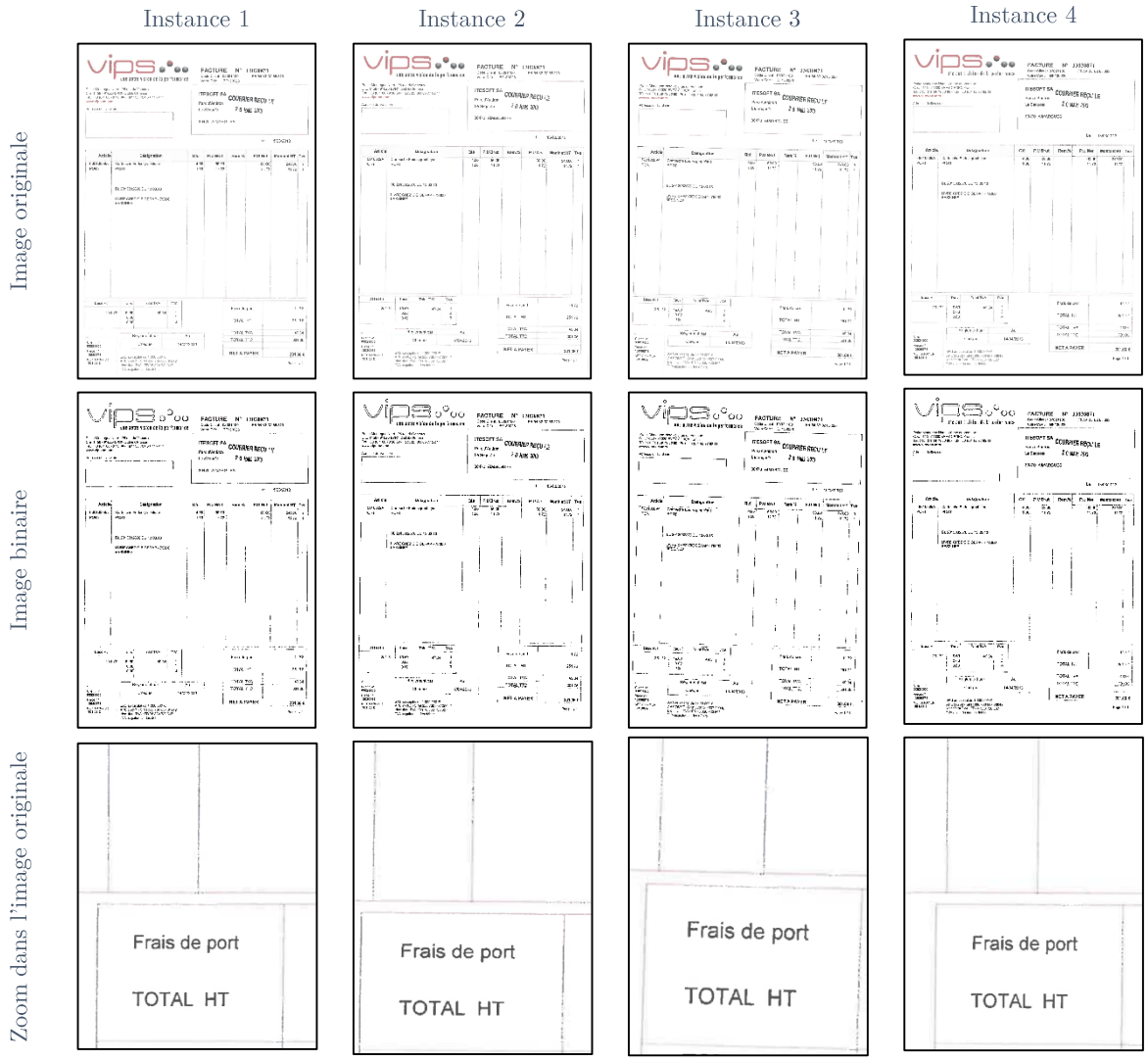
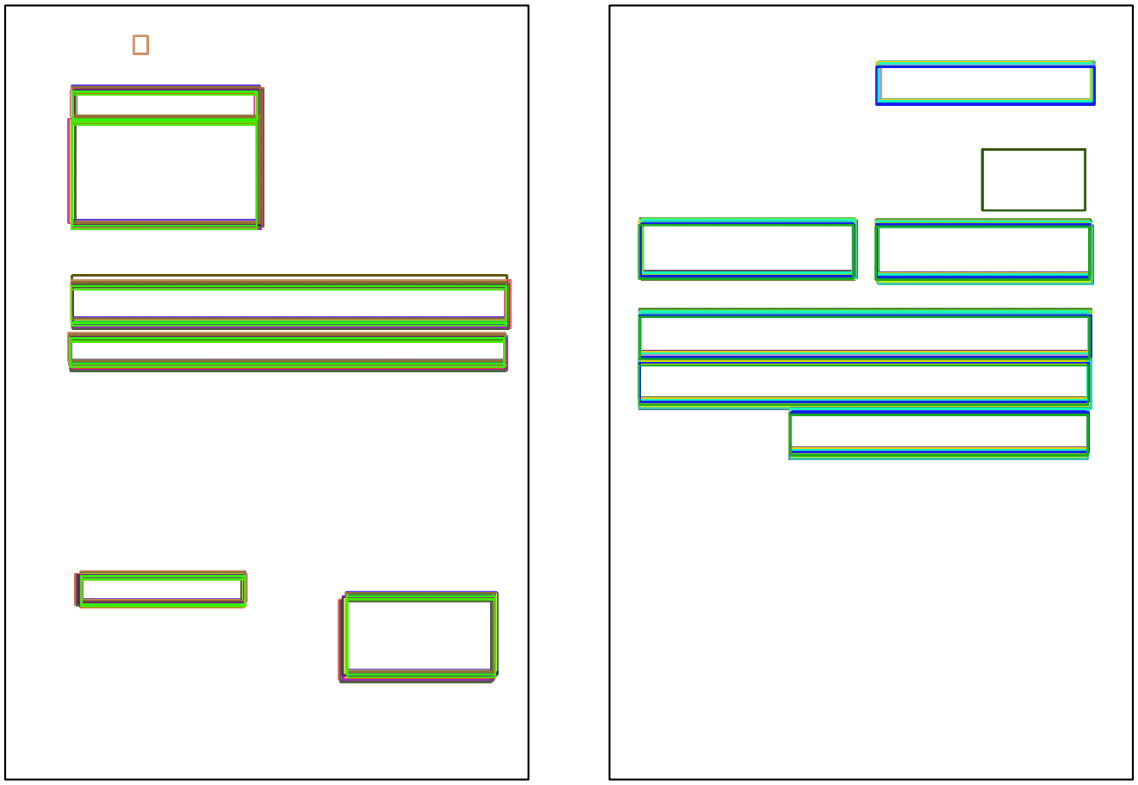


Figure 80 - Illustration du problème de l'extraction des traits sur quatre instances d'un document hybride.

Les résultats quantitatifs sont présentés dans le Tableau 8 où nous comparons nos résultats avec ceux obtenus en appliquant le logiciel de Tesseract [Smit09]. Nos résultats montrent une meilleure stabilité quelle que soit la mesure de score utilisée. Certains documents ne comportant pas de tableaux (document 9 et 10), obtiennent de mauvais scores, il s'agit de facturettes de supermarchés de mauvaise qualité où le logo a été détecté comme tableau mais cette erreur n'est pas stable dans l'ensemble des instances. La stabilité binaire nous permet de prendre en compte le biais lié au fait que les documents n'ont pas été recalés avant l'extraction des tableaux. On peut constater sur le Tableau 8 que cela conduit à un accroissement de l'évaluation des résultats dans la plupart des documents. Par contre, quand on passe d'un

risque à 0,7 à un risque à 0,8, plus contraignant, l'évaluation baisse, ce qui permet d'estimer le taux d'erreur spatiale associée aux tableaux détectés.



(a) (b)

Figure 81 - Exemples de superposition des résultats de notre méthode sur deux documents hybrides de « SETSTABLE » où chaque couleur représente une instance d'un document hybride. (a) Document hybride 1. (b) Document hybride 2.

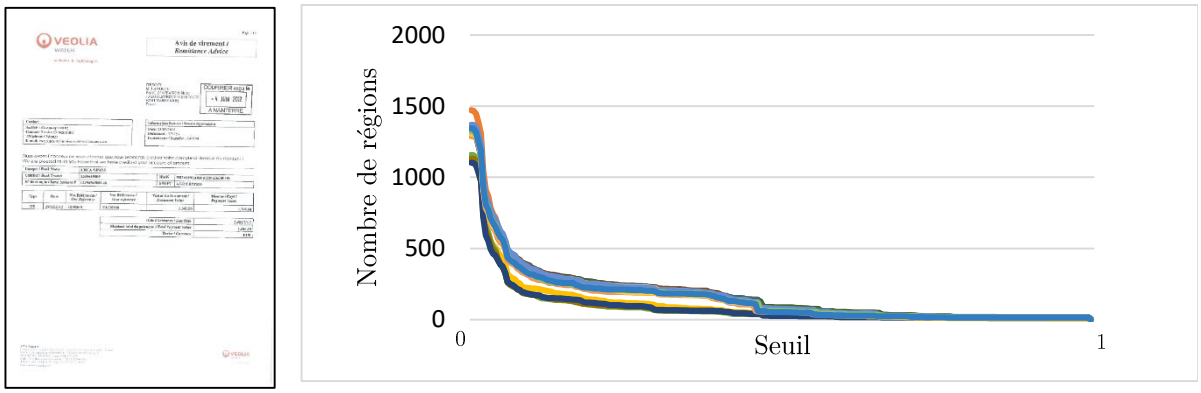
Tableau 8 - Évaluation de la stabilité de l'extraction de tableaux sur « SETSTABLE ».

Document hybride		1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
nombre d'occurrences		21	21	21	21	21	21	21	21	21	21	21	21	20	21	293
nombre de tableaux		6	6	1	6	2	0	0	3	0	0	0	7	0	0	31
Tesseract [Smit09]	SC (en %)	65	65	76	49	92	29	32	54	39	27	58	7	100	61	$\bar{x} = 65$
	$SC_{0.70}$ (en %)	52	57	77	28	90	13	10	47	27	7	45	2	100	63	$\bar{x} = 52$
	$SC_{0.80}$ (en %)	46	56	67	20	90	4	3	40	16	4	29	0	100	63	$\bar{x} = 46$
Méthode proposée	SC (en %)	80	80	87	82	86	100	100	88	59	57	81	49	100	73	$\bar{x} = 80$
	$SC_{0.70}$ (en %)	88	87	90	87	86	100	100	97	58	56	81	47	100	73	$\bar{x} = 88$
	$SC_{0.80}$ (en %)	62	73	84	78	74	100	100	91	57	56	81	36	100	73	$\bar{x} = 62$

4.6.2 Étude de la stabilité de la segmentation

La segmentation d'une page selon le media utilisé est la seconde étape du processus d'extraction de la mise en page que nous avons proposé. Cette segmentation est réalisée en étudiant le fond du document. En ne considérant que les grands segments présents dans le fond du document, des zones de fond, (cf. Section 3.3), ceux qui traversent presque entièrement le

document, nous améliorons la stabilité du processus global. En effet, ces zones constituent des segments plus stables que les segments plus petits. De leur détection dépend le nombre d'éléments de la segmentation de la page. Notre critère de sélection des séparateurs est leur longueur relative à la taille de la page. Nous avons illustré sur la Figure 82, le nombre de régions obtenues en fonction de la longueur des traits sélectionnés sur différentes instances (Figure 82 (b)) d'un document hybride (Figure 82 (a)). La stabilité est d'autant meilleure que le seuil est élevé.



(a) (b)

Figure 82 - Influence du seuil sur le nombre de régions. (a) Image Originale. (b) Graphique montrant le nombre de régions de différentes instances de document en fonction de la longueur des traits considérés.

Une étude expérimentale nous a conduit à considérer uniquement les segments possédant une longueur supérieure ou égale à 99 % de la taille du document. Cette valeur fournit en moyenne une image plus stable que les autres seuils. Malheureusement, les régions obtenues ne sont pas toujours pertinentes, comme nous pouvons le voir sur la Figure 83, où nous présentons trois segmentations par les plus grands segments (b, c et d) pour trois instances d'un document hybride (a). Les résultats possèdent encore quelques différences, notamment sur l'instance (c) ou une région verticale est apparue à cause de bruit présent sur l'image de l'instance. Ces régions doivent être encore divisées.

Ainsi comme nous l'avons montré dans le chapitre 3, les longs segments sont une aide stable à la segmentation mais leur détection ne suffit pas à assurer une segmentation stable du document.

(a) (b) (c) (d)

Figure 83 - Exemples des grands traits dans le fond d'un document hybride. (a) Image inverse binarisée. (b), (c) et (d) Résultats d'une segmentation par grands traits sur plusieurs instances du document hybride.

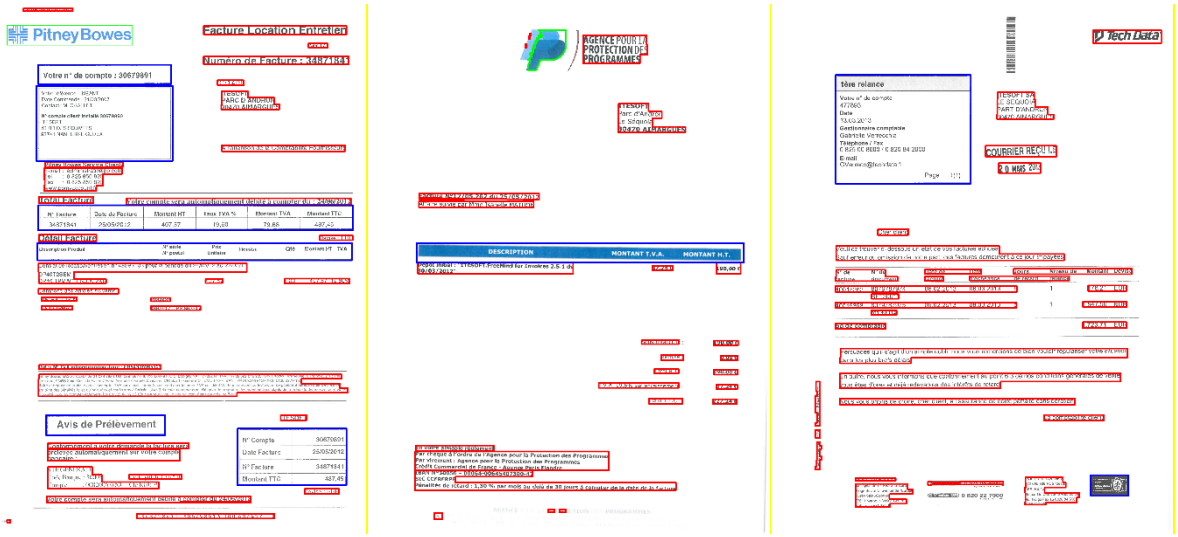
4.6.3 Étude de la stabilité de l'extraction de la mise en page

En appliquant la méthode décrite au chapitre 3 sur les images de la base « SETSTABLE » (cf. Figure 84), nous constatons une variété d'éléments présents dans les documents considérés, *eg.* tableaux, logos, textes. Il y a en particulier de nombreuses lignes de texte isolées. Dans les différentes instances, leurs différences de positionnement peuvent être négligeables. Nous réalisons deux estimations de la stabilité, l'une globalement en utilisant le DLD et l'autre en utilisant les scores de stabilité que nous avons proposés. Comme nous l'avons déjà remarqué cette dernière stratégie prend mieux en compte les formes des éléments et permet de mettre en évidence des problèmes dans la détection des régions (par exemple la non détection d'une colonne dans un tableau).

La Figure 85 illustre la nécessité d'un recalage entre les mises en page des instances d'un document hybride. L'étape de recalage se fait sur les images de résultats après avoir extrait la mise en page. Le Tableau 9 présente les résultats sur les images de la base « SETSTABLE » avec et sans recalage. De ce tableau nous pouvons observer que le recalage améliore les résultats, notamment sur le document 1 où nous avons présenté sur la Figure 85 le résultat du recalage. Sur ce document nous obtenons (en calculant le score de stabilité avec un seuil de 0,7) 82,29 % de stabilité contre 53,18 % sans le recalage. Le recalage améliore les résultats d'environ 10 % pour le score de stabilité globale et de 15% avec les seuils de 0,7 ou 0,8.

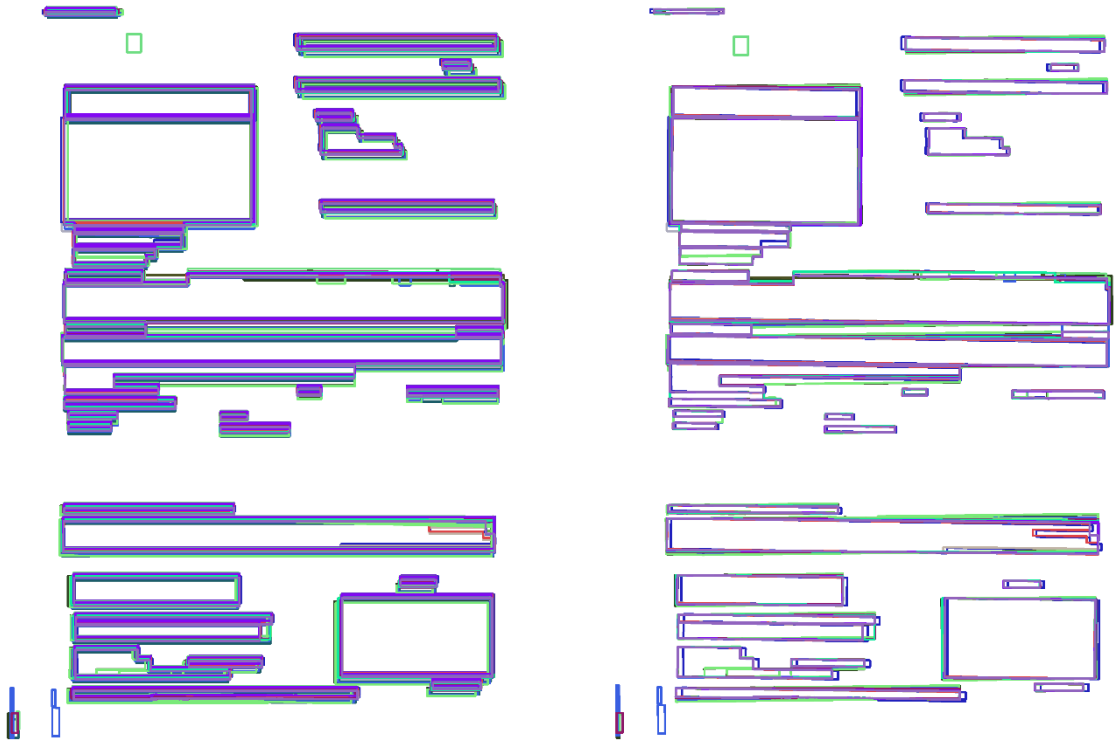
Les résultats les plus faibles sont encore ceux des tickets de caisse, des documents particulièrement difficiles à analyser. Ces résultats suggèrent qu'il est encore nécessaire d'améliorer la stabilité d'un tel système d'extraction de la mise en page pour l'utilisation dans

un contexte industriel. Nous allons maintenant voir le résultat de la stabilité grâce au descripteur de *layout* de Delaunay.



(a) (b) (c)

Figure 84 - Exemples de résultats de notre méthode pour l'extraction de la mise en page obtenue sur la base PRImA (légende couleur : bleu: tableau , rouge : texte et vert : image).



(a) (b)

Figure 85 - Exemples de superposition des résultats de notre méthode sur le document hybride 1 de « SETSTABLE » où chaque couleur représente une instance de ce document. (a) Sans recalage. (b) Avec recalage.

Tableau 9 - Évaluation de la stabilité de l'extraction de la mise en page sur « SETSTABLE ».

Document hybride	SC	SC (avec recalage)	$SC_{0.70}$	$SC_{0.70}$(avec recalage)	$SC_{0.80}$	$SC_{0.80}$(avec recalage)
1	63,65%	78,87%	53,18%	82,29%	41,27%	67,10%
2	67,80%	73,19%	62,49%	68,17%	53,53%	61,47%
3	77,50%	80,59%	74,64%	77,08%	62,21%	72,18%
4	67,83%	73,56%	58,08%	65,31%	47,74%	50,54%
5	71,41%	75,94%	67,79%	74,72%	54,88%	63,23%
6	48,89%	60,22%	31,79%	49,14%	24,86%	41,00%
7	52,48%	69,46%	36,56%	61,72%	26,12%	50,04%
8	65,51%	74,88%	55,82%	70,80%	43,21%	63,13%
9	51,53%	70,41%	36,20%	66,31%	27,53%	57,61%
10	52,48%	69,46%	36,56%	61,72%	26,12%	50,04%
11	61,23%	75,22%	48,14%	71,50%	35,82%	60,85%
12	69,71%	72,94%	62,04%	68,71%	53,38%	56,89%
14	59,80%	72,42%	48,89%	72,83%	37,96%	65,02%
Moyenne	62,29%	72,86%	51,70%	68,48%	41,13%	58,39%

Comme nous l'avions évoqué dans le Chapitre 3, nos méthodes sont étudiées pour des documents relativement droits, c'est-à-dire scannés sans inclinaison. Nous avons appliqué un prétraitement sur les documents afin de les redresser. Sur le dataset « SETSTABLE », nous avons évalué la correspondance des différents DLD obtenus avec et sans prétraitement pour les redresser (cf. Tableau 10). Comme nous l'avons vu dans le Chapitre 2, le choix des directions a rendu notre méthode sensible à l'inclinaison.

Tableau 10 - Résultats de la stabilité de l'extraction de la mise en page du document par la correspondance des résultats du DLD sur « SETSTABLE ».

Document Hybride	Sans pré-traitement	Avec pré-traitement
1	68,1 %	**99,52%**
2	34,3 %	**93,81%**
3	93,3 %	**100,00%**
4	85,2 %	**93,33%**
5	72,9 %	**98,10%**
6	**92,9 %**	81,90%
7	**91,0 %**	83,81%
8	61,9 %	**87,62%**
10	**87,6 %**	80,48%
11	73,8 %	**84,76%**
12	98,1 %	**100,00%**
13	84,2 %	NA
Total	78,6 %	**91,2%**

Les documents dont le score régresse sont des factures contenant de petits caractères de mauvaise qualité rendant difficile l'évaluation de l'angle d'inclinaison de la page. Nous observons que lorsque l'angle est correctement approximé, les résultats sont améliorés et deviennent très bons. C'est le cas des documents hybrides 3 et 12 qui passent à 100 % de correspondances entre les différents DLD. Les documents hybrides ayant les moins bons résultats correspondent aux tickets de caisse. Ce sont des documents très instables qui ont des problèmes d'orientation et la correction de l'inclinaison est plus complexe car de tels documents comportent peu d'informations. Un prétraitement plus complexe pourrait peut-être améliorer les résultats.

4.7 Synthèse et discussions

Dans ce chapitre, nous avons listé différentes méthodes de sécurisation existantes en fonction des différents cas de sécurisation considérés mais également en fonction de leurs contraintes pour finalement présenter la méthode de sécurisation proposée dans le projet SHADES dans lequel cette thèse s'inscrit. Cette méthode se base sur le contenu du document. Ainsi, le contenu doit être le même quelle que soit l'instance d'un document hybride. Cela sous-entend une stabilité de nos algorithmes d'extraction de la mise en page.

Ainsi, pour pouvoir estimer si notre méthode permet la sécurisation des documents hybrides comme nous l'avons déterminé dans le projet SHADES, nous avons défini les notions liées à la stabilité. Cela nous a permis de présenter différentes méthodes d'évaluation. Une étude de la stabilité des différentes méthodes présentées dans le Chapitre 3 a pu être effectuée selon ce critère.

Nos différentes méthodes, bien qu'ayant montré des résultats encourageants, ne permettent pas encore un niveau de confiance suffisant pour faire preuve devant un jury. En effet, notre méthode permet d'obtenir un taux de confiance quant au fait de savoir s'il s'agit du même document, mais en aucun cas une certitude. Cela permet toutefois de pouvoir créer une alerte qu'un être humain pourra vérifier. Dans l'ensemble, pour que le projet soit viable il faudrait améliorer la performance de nos méthodes. Comme nous avons pu l'observer, la binarisation, qui est une des étapes fondamentales de notre méthode, n'est pas stable pour les différentes instances d'un même document. Nous pourrons ainsi améliorer les résultats en considérant les documents en niveaux de gris ou en couleurs (en nous affranchissant alors d'une étape de binarisation) ou en améliorant par des prétraitements la qualité des images de documents.

Conclusion

Bilan et contributions

Les travaux réalisés durant cette thèse s'inscrivent dans l'analyse et le traitement des images de documents, ainsi que dans la caractérisation des images. En particulier des méthodes ont été développées dans le but d'analyser la mise en page pour sécuriser les documents. Pour différentes instances de documents hybrides, des résultats identiques par rapport au nombre de régions et aux positions relatives des différentes régions les unes par rapport aux autres permettent de savoir qu'on considère le même document.

Dans la première partie de cette thèse, nous avons proposé une nouvelle représentation pour caractériser les images en considérant que celles-ci ne sont plus composées de primitives pixels mais d'un ensemble de segments ayant des directions et des longueurs différentes. Pour cela, nous avons défini de nouvelles transformées fondées sur les segments dans les images. Ces transformées ont été présentées en détail par leurs définitions, leurs propriétés, leurs implémentations, *etc.* Nous avons également présenté quelques cas d'applications. Elles permettent notamment, en se basant sur la dualité fond / forme, d'extraire des caractéristiques. Ces caractéristiques permettant à leur tour d'analyser les images.

Cette analyse nous a permis, dans un premier temps, d'extraire les séparateurs et les tableaux matérialisés. L'extraction de ces éléments en particulier, avant de s'intéresser aux autres et de continuer le traitement, est stratégique. En effet, ces éléments ont des comportements particuliers : premièrement, ils perturbent l'extraction du reste de la mise en page et deuxièmement, ce sont des éléments extrêmement instables car ils sont composés de

traits fins qui disparaissent généralement de manière aléatoire lors de l'impression et de la numérisation. Une analyse sur le fond, par ces transformées, nous permet également de segmenter le document. La labélisation des régions est également réalisée par l'analyse des segments qui composent le document. La méthode d'extraction de la mise en page est ainsi réalisée sans utiliser de méthode d'apprentissage supervisé. Des évaluations ont été ensuite réalisées pour évaluer les différentes étapes qui composent notre méthode. Ces évaluations ont été réalisées sous l'angle de la « qualité » des résultats. Nous cherchons à savoir si la mise en page a été correctement trouvée mais ce n'est pas la seule façon d'évaluer et dans le cadre du volet sécurité, le besoin d'un autre type d'évaluation a été mis en évidence.

La deuxième partie de ces travaux concerne donc le volet sécurité. Le projet SHADES dans lequel s'inscrit cette thèse, a pour objectif de sécuriser les documents dits « hybrides ». Dans ce contexte, nous avons présenté différentes méthodes permettant de sécuriser les documents qu'ils soient matériels ou immatériels. Après avoir rappelé le principe du projet, nous avons montré le lien entre la stabilité des résultats sur les instances d'un même document et la sécurité dudit document. Nous avons proposé des définitions pour la stabilité et les termes connexes tels qu'égalité, robustesse et sensibilité. Différentes méthodes pour évaluer celles-ci ont été présentées pour finalement analyser les résultats de nos méthodes sous cet angle.

Les travaux issus de cette thèse ont à ce jour conduit à la publication dans des conférences internationales [ACKO17] et [AACK19].

Perspectives de recherche

À l'issue des travaux réalisés dans cette thèse, nous avons pu ouvrir plusieurs pistes de réflexion. Comme nous l'évoquions dans le Chapitre 2, les transformées sont calculées sur des images binaires. Or, la binarisation peut engendrer des pertes d'informations plus ou moins importantes, les transformées ont alors été étendues aux niveaux de gris. La principale difficulté que nous voyons à présent concerne la distinction entre le fond et la forme. Comme nous l'avions expliqué dans la Section 2.8, les segments n'étant plus définis que par la longueur ou l'orientation, on ne peut plus savoir à quoi correspondent les différents éléments qui composent celui-ci. Dans un premier temps, nous pouvons utiliser des informations grâce à l'utilisation d'une image binaire. Mais il serait intéressant de rajouter de l'information se basant sur la couleur, entre autres, pour faire une analyse plus complète.

D'autres applications peuvent également être réalisées grâce à l'analyse des traits présents dans les images et extraits par nos transformées. Nous avions notamment commencé par analyser la segmentation des lignes dans des documents manuscrits par analyse des

interlignes trouvées en appliquant les transformées sur le fond du document. Ici, l'un des problèmes concerne les cheminées présentes. Celles-ci perturbent le traitement, ne permettent plus de reconnaître l'importance d'un trait présent dans le fond, les confondant ainsi avec des traits séparateurs. Mais, en les considérant, nous pouvons obtenir différents morceaux de lignes qu'un post-traitement devra lier. Le problème revient ici à traiter l'ensemble des morceaux pour créer de véritables lignes. Pour associer ces différents morceaux, nous avons pensé à modifier un algorithme génétique pour l'adapter à notre problème, la difficulté étant de trouver la bonne fonction de coût. Une autre approche utilise les graphes. Nous pouvons ainsi créer un arbre considérant chaque morceau comme un nœud et la distance entre les éléments comme attribut des arêtes. Pour trouver les lignes, il faut ainsi trouver différents chemins dans le graphe. La difficulté est ici de trouver la bonne mesure de distance ayant un coût élevé si les deux morceaux de lignes ne sont pas sur la même ligne et faible si deux extrémités opposées sont proches. L'application de l'ACP (l'analyse en composantes principales) sur les coordonnées des pixels d'un morceau de lignes permet de connaître l'axe dans lequel le morceau est orienté. En utilisant cette information, nous pouvons adapter la mesure de distance, ce qui permettrait de réduire le coût lorsque l'on se déplace sur cet axe et l'augmenter lorsque l'on diverge de cet axe.

Parmi les perspectives liées à l'application finale, qui consiste à sécuriser les documents hybrides, nous pouvons améliorer les résultats en matière de stabilité. Plusieurs manières peuvent être envisagées. Ainsi, nous pouvons prendre en compte plus d'informations par l'utilisation de transformées en niveaux de gris ou en couleur, mais également utiliser de meilleurs prétraitements avant de lancer les opérations. La dernière solution est à notre avis recommandable, notamment dans le cadre de tickets de caisse ou de facturettes car l'encre et les conditions de stockage ne sont généralement pas assez pérennes.

Bibliographie

[AACK19] H. Alhéritière, W. Amaïeur, F. Cloppet, C. Kurtz, J.-M. Ogier, N. Vincent, Straight Line Reconstruction for Fully Materialized Table Extraction in Degraded Document Images, in *Discrete Geometry for Computer Imagery (DGCI)*, 2019: pp. 317–329.

[AbBö16] S. Abramova, R. Böhme, Detecting Copy – Move Forgeries in Scanned Text Documents, in *Society for Imaging Science and Technology*, 2016: pp. 1–9.

[ACKO17] H. Alhéritière, F. Cloppet, C. Kurtz, J. Ogier, N. Vincent, A document straight line based segmentation for complex layout extraction, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2017: pp. 1126–1131.

[ACPP11] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher, Historical Document Layout Analysis Competition, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2011: pp. 1516–1520.

[ACPP13] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher, ICDAR 2013 Competition on Historical Book Recognition (HBR 2013), in *International Conference on Document Analysis and Recognition (ICDAR)*, 2013: pp. 1459–1463.

[ACPP15] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher, Competition on Recognition of Documents with Complex Layouts –, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2015: pp. 1151–1155.

[AkTo11] C. Akinlar, C. Topal, EDLines : A real-time line segment detector with a false detection control, *Pattern Recognition Letters*. vol. 32 pp. 1633–1642 (2011).

[AnGB05] A. Antonacopoulos, B. Gatos, D. Bridson, ICDAR2005 Page Segmentation Competition The dataset, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2005: pp. 75–79.

[AnGB07] A. Antonacopoulos, B. Gatos, D. Bridson, ICDAR2007 Page Segmentation Competition, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2007: pp. 1279–1283.

[AnGK03] A. Antonacopoulos, B. Gatos, D. Karatzas, ICDAR 2003 Page Segmentation Competition The dataset, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2003: pp. 688–692.

[Anto98] A. Antonacopoulos, Page segmentation using the description of the background, *Computer Vision and Image Understanding*. vol. 70 pp. 350–369 (1998).

[APBP09] A. Antonacopoulos, S. Pletschacher, D. Bridson, C. Papadopoulos, ICDAR2009 Page Segmentation Competition, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2009: pp. 1370–1374.

[Arta19] C. Artaud, Détection des fraudes : de l'image à la sémantique du contenu, Thèse

de doctorat. Université de La Rochelle, 2019.

[ASDO18] C. Artaud, N. Sidère, A. Doucet, J. Ogier, V. Poulain d'Andecy, Find it ! Fraud Detection Contest Report, in *International Conference on Pattern Recognition (ICPR)*, 2018: pp. 13–18.

[ASSL12] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 34 pp. 2274–2281 (2012).

[Auge13] O. Augereau, Reconnaissance et classification d'images de documents, Thèse de Doctorat. Université Bordeaux 1, 2013.

[Ball81] D.H. Ballard, Generalizing the hough transform to detect arbitrary shapes, *Pattern Recognition (PR)*. vol. 13 pp. 111–122 (1981).

[BaMo17] E. Barker, N. Mouha, Recommendation for Triple Data Encryption Algorithm (TDEA) Block Cipher. Research report, 2017.

[Bark16] E. Barker, Guideline for using cryptographic standards in the federal government : cryptographic mechanisms. Research report, National Institute of Standards and Technology, 2016.

[BASB10] S.S. Bukhari, M.I.A. Al Azawi, F. Shafait, T.M. Breuel, Document image segmentation using discriminative learning over connected components, in *International Workshop on Document Analysis Systems (DAS)*, 2010: pp. 183–190.

[BGKW95] H. Breu, J. Gil, D. Kirkpatrick, M. Werman, Linear Time Euclidean Distance Transform Algorithms 1 Introduction, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 17 pp. 529–533 (1995).

[BGTF13] R. Bertrand, P. Gomez-krämer, O.R. Terrades, P. Franco, J. Ogier, A System Based On Intrinsic Features for Fraudulent Document Detection, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2013: pp. 106–110.

[Bloo91] D.S. Bloomberg, Multiresolution Morphological Approach to Document Image Analysis, in *International Conference on Document Analysis and Recognition (ICDAR)*, 1991: pp. 963–971.

[BMAC07] H.S. Baird, M. a Moll, C. An, M.R. Casey, Document Image Content Inventories, in *Document Recognition and Retrieval (DRR)*, 2007.

[BoBB13] M.-R. Bouguelia, Y. Belaid, A. Belaïd, Document image and zone classification through incremental learning, in *International Conference on Image Processing (ICIP)*, 2013: pp. 4230–4234.

[Bres65] J.E. Bresenham, Algorithm for computer control of a digital plotter, *IBM Systems Journal*. vol. 4 pp. 25–30 (1965).

[BSND18] S. Bhowmik, R. Sarkar, M. Nasipuri, D. Doermann, Text and non-text separation in offline document images: a survey, *International Journal on Document Analysis and*

Recognition (IJDAR). vol. 21 pp. 1–20 (2018).

[BVMR15] B. V. Bharath, A.S. Vilas, K. Manikantan, S. Ramachandran, Iris recognition using radon transform thresholding based feature extraction with gradient-based isolation as a pre-processing technique, in *International Conference on Industrial and Information Systems (ICIIS)*, 2015: pp. 1–8.

[CaCa08] L. Caponetti, C. Castiello, Document page segmentation using neuro-fuzzy approach, *Applied Soft Computing*. vol. 8 pp. 118–126 (2008).

[CAKE13] R. Cohen, A. Asi, K. Kedem, J. El-Sana, I. Dinstein, Robust text and drawing segmentation algorithm for historical documents, in *International Workshop on Historical Document Imaging and Processing (HIP)*, 2013: p. 110.

[Cann86] J. Canny, A Computational Approach to Edge Detection, *Pattern Analysis and Machine Intelligence (PAMI)*. vol. 8 pp. 679–698 (1986).

[CaTV17] J. Calvo-zaragoza, A.H. Toselli, E. Vidal, Handwritten Music Recognition for Mensural Notation : Formulation , Data and Baseline Results, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2017: pp. 1081–1086.

[CFGM97] F. Cesarini, E. Francesconi, M. Gori, S. Marinai, J.Q. Sheng, G. Soda, A neural-based architecture for spot-noisy logo recognition, in *International Conference on Document Analysis and Recognition (ICDAR)*, 1997.

[ChYL13] K. Chen, F. Yin, C.L. Liu, Hybrid page segmentation with efficient whitespace rectangles extraction and grouping, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2013: pp. 958–962.

[CiSc01] G. Ciocca, R. Schettini, Content-based similarity retrieval of trademarks using relevance feedback, *Pattern Recognition (PR)*. vol. 34 pp. 1639–1655 (2001).

[ClAP17] C. Clausner, A. Antonacopoulos, S. Pletschacher, ICDAR2017 Competition on Recognition of Documents with Complex Layouts – RDCL2017, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2017: pp. 1404–1410.

[CMSS02] F. Cesarini, S. Marinai, L. Sarti, G. Soda, Trainable table location in document images, in *International Conference on Pattern Recognition (ICPR)*, 2002: pp. 236--240.

[CoBr14] M. Cote, A. Branzan Albu, Texture sparseness for pixel classification of business document images, in *International Journal on Document Analysis and Recognition (IJDAR)*, 2014: pp. 257–273.

[CoGC14] D. Coppi, C. Grana, R. Cucchiara, Illustrations Segmentation in Digitized Documents Using Local Correlation Features, in *Italian Research Conferance on Digital Libraries (IRCDL)*, 2014: pp. 76–83.

[CoPV15] D. Cozzolino, G. Poggi, L. Verdoliva, Efficient dense-field copy-move forgery detection, *IEEE Transactions on Information Forensics and Security*. vol. 10 pp. 2284 (2015).

[CSHI17] K. Chen, M. Seuret, J. Hennebert, R. Ingold, Convolutional Neural Networks for Page Segmentation of Historical Document Images, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2017: pp. 965–970.

[DaRi02] J. Daemen, V. Rijmen, The Design of Rijndael: AES - The Advanced Encryption Standard. Research report, 2002.

[Debl95] I. Debled-Rennesson, Etude et reconnaissance des droites et plans discrets, Thèse de Doctorat. Université de Strasbourg, 1995.

[DeMM00] A. Desolneux, L. Moisan, J. Morel, Meaningful Alignments, *International Journal of Computer Vision*. vol. 40 pp. 7–23 (2000).

[DoRW96] D. Doermann, E. Rivlin, I. Weiss, Applying algebraic and differential invariants for logo recognition, *Machine Vision and Applications*. vol. 9 pp. 73–86 (1996).

[EsGO15] S. Eskenazi, P. Gomez-krämer, J. Ogier, The Delaunay document layout descriptor, in *Symposium on Document Engineering (DocEng)*, 2015: pp. 167–175.

[Eske17] S. Eskenazi, On the stability of document analysis algorithms : application to hybrid document hashing technologies, Thèse de Doctorat. Université de La Rochelle, 2017.

[FCTB08] R. Fabbri, L.D.F. Costa, J.C. Torelli, O.M. Bruno, 2D Euclidean distance transform algorithms, *ACM Computing Surveys*. vol. 40 pp. 1–44 (2008).

[FeTS14] M. Felhi, S. Tabbone, M.V. ortiz Segovia, Multiscale Stroke-Based Page Segmentation Approach, in *International Workshop on Document Analysis Systems (DAS)*, 2014: pp. 6–10.

[FrGo00] J. Fridrich, M. Goljan, Robust Hash Functions for Digital Watermarking, in *International Conference on Information Technology: Coding and Computing*, 2000: pp. 178–183.

[GDPP05] B.G. Gatos, D. Danatsas, I. Pratikakis, S.J. Perantonis, Automatic table detection in document images, in *International Conference on Pattern Recognition and Data Mining (ICPRDM)*, 2005: pp. 609–618.

[GhBe16] N. Ghanmi, A. Belaïd, Recognition-based Approach of Numeral Extraction in Handwritten Chemistry Documents using Contextual Knowledge, in *Workshop on Document Analysis Systems*, 2016: pp. 251–256.

[GHOO13] M. Gobel, T. Hassan, E. Oro, G. Orsi, ICDAR 2013 table competition, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2013: pp. 1449–1453.

[GLEE06] D. Gaceb, F. Lebourgeois, V. Eglin, H. Emptoz, Contribution to the Automatic Recognition of Business Documents, in *International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, 2006.

[GPKB18] P. Gomez-krämer, A.T. Phan ho, W. Khlif, J. Burie, N. Sidere, Comparison

process between two documents and the localisation of differences. Research report, 2018.

[GQMS17] A. Gilani, S.R. Qasim, I. Malik, F. Shafait, Table Detection Using Deep Learning, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2017: pp. 771–776.

[GrRB12] D. Grzejszczak, Y. Rangoni, A. Belaïd, Séparation manuscrit et imprimé dans des documents administratifs complexes par utilisation de SVM et regroupement, in *Colloque International Francophone Sur l'Ecrit et Le Document (CIFED)*, 2012.

[GuVZ95] N. Guil, J. Villalba, E.L. Zapata, A Fast Transform for Segment Detection, *IEEE Transactions on Image Processing*. vol. 4 pp. 1541–1548 (1995).

[GYLJ17] L. Gao, X. Yi, Y. Liao, Z. Jiang, Z. Yan, Z. Tang, A Deep Learning-based Formula Detection Method for PDF Documents, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2017: pp. 553–558.

[HaCV14] S. Hamrouni, F. Cloppet, N. Vincent, Séparation imprimé / manuscrit par étude de la linéarité et de la régularité du texte, in *Colloque International Francophone Sur l'Ecrit et Le Document (CIFED)*, 2014: pp. 155–170.

[HaHP95] J.H.J. Ha, R.M. Haralick, I.T. Phillips, Recursive X-Y cut using bounding boxes of connected components, in *International Conference on Document Analysis and Recognition (ICDAR)*, 1995: pp. 952–955.

[Hall98] M.A. Hall, Correlation-based Feature Selection for Machine Learning, Thèse de Doctorat. The University of Waikato, 1998.

[HBSM08] Z. Haddad, A. Beghdadi, A. Serir, A. Mokraoui, Fingerprint identification using radon transform, in *International Workshops on Image Processing Theory, Tools and Applications (IPTA)*, 2008: pp. 1–7.

[HCTM12] H. He, F. Chen, H. Tai, S. Member, T. Kalker, J. Zhang, Performance Analysis of a Block-Neighborhood- Based Self-Recovery Fragile Watermarking Scheme, *IEEE Transactions on Information Forensics and Security*. vol. 7 pp. 185–196 (2012).

[HGYT16] L. Hao, L. Gao, X. Yi, Z. Tang, A Table Detection Method for PDF Documents Based on Convolutional Neural Networks, in *International Workshop on Document Analysis Systems (DAS)*, 2016: pp. 287–292.

[HLHN15] T.K. Huynh, T. Le-Tien, K. V. Huyunh, S.C. Nguyen, A Survey on Image Forgery Detection Techniques, in *International Conference on Computing & Communication Technologies Research, Innovation, and Vision for Future (RIVF)*, 2015: pp. 71–76.

[Houg62] P.V. Hough, Method and means for recognizing complex patterns, (1962).

[JaDo12] R. Jain, D. Doermann, Logo retrieval in document images, in *International Workshop on Document Analysis Systems (DAS)*, 2012: pp. 135–139.

[JaVa98] A.K. Jain, A. Vailaya, Shape-based retriecal: a case study with trademar image databases, *Pattern Recognition (PR)*. vol. 31 pp. 1369–1390 (1998).

[JMER07] N. Journet, R. Mullot, V. Eglin, J.-Y. Ramel, Analyse d'Images de Documents Anciens: une Approche Texture, *Traitement Du Signal*. vol. 24 pp. 461–479 (2007).

[KeSB06] D. Keysers, F. Shafait, T.M. Breuel, Document image zone classification, in *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2006.

[KiDe01] T. Kieninger, A. Dengel, Applying the T-Recs Table Recognition System to the Business Letter Domain, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2001: pp. 518–522.

[KiKi97] Y.-S. Kim, W.-Y. Kim, Content-based trademark retrieval system using visually salient features, in *Computer Society Conference on Computer Vision and Pattern Recognition*, 1997: pp. 307–312.

[KiSI98] K. Kise, A. Sato, M. Iwata, Segmentation of Page Images Using the Area Voronoi Diagram, *Computer Vision and Image Understanding*. vol. 70 pp. 370–382 (1998).

[Koff35] K. Koffka, Principles of gestalt psychology, (1935).

[Koru17] P. Korus, Digital image integrity – a survey of protection and verification techniques, *Digital Signal Processing*. vol. 71 pp. 1–26 (2017).

[KSFV09] K. Khurshid, I. Siddiqi, C. Faure, N. Vincent, Comparison of Niblack inspired binarization methods for ancient documents, in *Document Recognition and Retrieval (DRR)*, 2009.

[LaMB06] C.H. Lampert, L. Mei, T.M. Breuel, Printing Technique Classification for Document Counterfeit Detection, in *International Conference on Computational Intelligence and Security*, 2006: pp. 639–644.

[LiLM86] H. Li, M.A. Lavin, R.J.L.E. Master, Fast Hough Transform : A Hierarchical Approach, in *Computer Vision, Graphics, and Image Processing*, 1986: pp. 139–161.

[LoBr98] willizm S. Lovegrove, D.F. BrailFord, Document analysis of PDF files : methods , results and implications, *Electronic Publishing*. vol. 8 pp. 207–220 (1998).

[LVTO12] V.P. Le, M. Visani, C. De Tran, J. Ogier, Logo Spotting For Document Categorization, in *International Conference on Pattern Recognition (ICPR)*, 2012: pp. 3484–3487.

[MaMS05] S. Marinai, E. Marino, G. Soda, Layout based document image retrieval by means of XY tree reduction, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2005: pp. 432–436.

[MCDC06] S. Mandal, S.P. Chowdhury, a K. Das, B. Chanda, A simple and effective table detection system from document images, *International Journal on Document Analysis and Recognition (IJDAR)*. vol. 8 pp. 172–182 (2006).

[MeGB17] L. Melinda, R. Ghanapuram, C. Bhagvati, Document Layout Analysis using Multigaussian Fitting, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2017: pp. 747–752.

[MeKW97] B.M. Mehtre, M.S. Kankanhalli, Wing Foon Lee, Shape measures for content based image retrieval: A comparison, *Information Processing & Management*. vol. 33 pp. 319–337 (1997).

[Meun05] J.L. Meunier, Optimized XY-cut for determining a page reading order, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2005: pp. 347–351.

[MuCh15] P. Mukhopadhyay, B.B. Chaudhuri, A survey of Hough Transform, *Pattern Recognition*. vol. 48 pp. 993–1010 (2015).

[NaSe84] G. Nagy, S.C. Seth, Hierarchical Representation of Opically Scanned Documents, in *International Conference on Pattern Recognition (ICPR)*, 1984: pp. 347–349.

[NaTZ12] N. Nacereddine, S. Tabbone, D. Ziou, Object Recognition Using Radon Transform-Based RST Parameter Estimation Transform-Based RST Parameter Estimation, in *Advanced Concepts for Intelligent Vision Systems (ACIVS)*, 2012: pp. 515–526.

[NTBT16] M. Nouredanesh, H.R. Tizhoosh, E. Banijamali, J. Tung, Radon-Gabor barcodes for medical image retrieval, in *International Conference on Pattern Recognition (ICPR)*, 2016: pp. 1309–1314.

[Otsu79] N. Otsu, A threshold selection method from gray-level histograms, *IEEE Transactions on Systems, Man, and Cybernetics*. vol. 9 pp. 62–66 (1979).

[PhZh91] I.T. Phillips, J. Zhou, Page Segmentation by White Stream, in *International Conference on Document Analysis and Recognition (ICDAR)*, 1991: pp. 945–953.

[Piva13] A. Piva, An Overview on Image Forensics, *ISRN Signal Processing*. vol. 8 pp. 14–25 (2013).

[Rado17] J. Radon, über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten, *Classic Papers in Modern Diagnostic Radiology*. (1917).

[Ragn93] I. Ragnemalm, The Euclidean distance transformation in arbitrary dimensions, *Pattern Recognition Letters*. vol. 14 pp. 883–888 (1993).

[RCVF03] J.-Y. Ramel, M. Crucianu, N. Vincent, C. Faure, Detection, extraction and representation of tables, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2003: pp. 374–378.

[RoEH11] H. Rojbani, I. Elouedi, A. Hamouda, R-signature: A new signature based on Radon Transform and its application in buildings extraction, in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2011: pp. 490–495.

[ShSm10] F. Shafait, R. Smith, Table detection in heterogeneous documents, in *International Workshop on Document Analysis Systems (DAS)*, 2010: pp. 65–72.

[Smit09] R. Smith, Hybrid page layout analysis via tab-stop detection, in *International*

Conference on Document Analysis and Recognition (ICDAR), 2009: pp. 241–245.

[SoSa98] A. Soffer, H. Samet, Using negative shape features for logo similarity matching, in *International Conference on Pattern Recognition (ICPR)*, 1998.

[Suth10] P. Sutheebanjard, A Modified Recursive X-Y Cut Algorithm for Solving Block Ordering Problems, in *International Conference on Computer Engineering and Technology*, 2010: pp. 307–311.

[Trup05] É. Trupin, La reconnaissance d 'images de documents : Un panorama Document images recognition : A survey, *Traitement Du Signal*. vol. 22 pp. 159–190 (2005).

[TZCM16] J. Thies, M. Zollh, S. Christian, T. Matthias, Face2Face : Real-time Face Capture and Reenactment of RGB Videos, in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016: pp. 2387–2395.

[VaBe12] J. Vauthier, A. Belaïd, Segmentation et classification des zones d'une page de document, in *Colloque International Francophone Sur l'Ecrit et Le Document (CIFED)*, 2012.

[ViDo12] R. Vieux, J. Domenger, Hierarchical Clustering Model for Pixel-Based Classification of Document Images, in *International Conference on Pattern Recognition (ICPR)*, 2012: pp. 290–293.

[WaPH02] Y. Wang, I.T. Phillips, R.M. Haralick, A method for document zone content classification, *Object Recognition Supported by User Interaction for Service Robots*. vol. 3 pp. 0–3 (2002).

[WaPH06] Y. Wang, I.T. Phillips, R.M. Haralick, Document zone content classification and its performance evaluation, *Pattern Recognition (PR)*. vol. 39 pp. 57–73 (2006).

[WaWC82] F.M. Wahl, K.Y. Wong, R.G. Casey, Block segmentation and text extraction in mixed text/image documents, *Computer Graphics and Image Processing*. vol. 19 pp. 94 (1982).

[WiAS11] A. Winder, T. Andersen, E.H.B. Smith, Extending page segmentation algorithms for mixed-layout document processing, in *International Conference on Document Analysis and Recognition (ICDAR)*, 2011: pp. 1245–1249.

[XuOK90] L. Xu, E. Oja, P. Kultanen, A new curve detection method: randomized Hough transform (RHT), *Pattern Recognition Letters*. vol. 11 pp. 331–338 (1990).

[ZaBC02] R. Zanibbi, D. Blostein, J.R. Cordy, Recognizing mathematical expressions using tree transformation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 24 pp. 1455–1467 (2002).

[ZaBC04] R. Zanibbi, D. Blostein, J.R. Cordy, A survey of table recognition: Models, observations, transformations, and inferences, *International Journal on Document Analysis and Recognition (IJDAR)*. vol. 7 pp. 1–16 (2004).

[ZENM13] F. Zirari, A. Ennaji, S. Nicolas, D. Mammass, A simple text/graphic separation

method for document image segmentation, in *International Conference on Computer Systems and Applications (AICCSA)*, 2013: pp. 1–4.

[ZhCo07] Q. Zhang, I. Couloigner, Accurate centerline detection and line width estimation of thick lines using the radon transform, *IEEE Transactions on Image Processing*. vol. 16 pp. 310–316 (2007).

[ZhWa07] X. Zhang, S. Wang, Statistical Fragile Watermarking Capable of Locating Individual Tampered Pixels, *Signal Processing Letters*. vol. 14 pp. 727–730 (2007).

[ZJYW17] X. Zhu, Y. Jiang, S. Yang, X. Wang, W. Li, P. Fu, H. Wang, Z. Luo, Deep Residual Text Detection Network for Scene Text, pp. 807–812 (2017).