



HAL
open science

Etat de l'art sur des outils logiciels, actuels ou en développement et des labos et équipes de recherche

Nicolas Sidère

► **To cite this version:**

Nicolas Sidère. Etat de l'art sur des outils logiciels, actuels ou en développement et des labos et équipes de recherche. [Rapport Technique] BnF. 2018. hal-02432577

HAL Id: hal-02432577

<https://bnf.hal.science/hal-02432577>

Submitted on 8 Jan 2020

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

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

Etat de l'art sur des outils logiciels, actuels ou en développement et des labos et équipes de recherche

VALCONUM

Université de La Rochelle – Faculté des Sciences et Technologies – L3i

Bât. Pascal, Av Michel Crépeau

17042 La Rochelle Cedex 1

Tél. : 05 46 45 72 20 – nelly.chauvin@valconum.org

Œuvre collective menée sous la direction de Nicolas Sidère

Introduction	3
Les contours du domaine multimédia et de cet état de l'art	3
Présentation de la BnF	4
Les corpus images de la BnF	5
Architecture général d'un système de recherche	8
Synthèse	11
Pré-traitement et segmentation (Extraction de la structure non labellisée)	12
2.1 Pré-traitement	12
2.2 Segmentation ou extraction du layout physique	15
Indexation sémantique des images	21
Indexation par le contenu visuel	29
4.1 Introduction	29
4.2 Caractéristiques	29
Approches par description globale	30
Approches par description spatiale	30
Approches par description locale	32
4.3 Apprentissage des caractéristiques	34
Logiciels et boîtes à outils disponibles	35
Services commerciaux	35
Frameworks et réseaux pré-entraînés pour la classification	38
Frameworks classiques	38
Modèles pré-entraînés : les zoos	40
Logiciels d'indexation pour la recherche d'images similaires	41
Principaux acteurs	42
Conclusion de l'étude.	47
Références Bibliographiques	50

Ce document vise à dresser un état de l'art de l'indexation de contenus multimédias à l'intention de la Bibliothèque Nationale de France (BnF), mettant en lumière les techniques actuelles et les logiciels et outils afférents et discutant de leurs limites. Il a été produit à la demande de la BnF en partenariat avec les membres des laboratoires suivants (par ordre alphabétique) : CVC (Barcelone), IRISA (Rennes), LIPADE (Paris V), L3i (La Rochelle). On dressera tout d'abord un aperçu général du domaine de l'indexation multimédia, avant de discuter des grandes familles de techniques d'indexation appliquées aux images et de présenter les outils et logiciels disponibles permettant une indexation des données.

1. Introduction

a. Les contours du domaine multimédia et de cet état de l'art

Les contenus multimédias mêlent plusieurs modalités afin de communiquer un message ou une information à une ou plusieurs personnes. Les modalités traditionnelles sur lesquelles s'appuie le domaine du multimédia sont le texte, les images, la vidéo et l'audio. On voit dans cette définition se profiler une notion importante : celle d'utilisateur, i.e., une personne cherchant des informations, ou simplement un divertissement, dans la consommation de contenus multimédias et à qui le processus d'indexation est dirigé.

L'indexation multimédia quant à elle peut être définie comme un ensemble de techniques permettant de décrire, indexer et retrouver, si possible de manière efficace, des contenus multimédias référencés dans une base de contenus [Gros 2007]. La recherche de contenus est souvent considérée par le biais de moteurs de recherche, parfois abusivement appelés moteurs d'indexation, permettant de retrouver des contenus pertinents à partir d'une requête. Dans le domaine du multimédia, cette dernière peut prendre de nombreuses formes que nous décrirons par la suite. On peut par exemple rechercher les contenus parlant d'un sujet ou d'une personne, les textes contenant un ensemble de mots, les images dans lesquels on voit un objet, ou encore les textes et/ou images proche d'un texte et/ou d'une image donnée.

Pour des raisons historiques, le domaine de l'indexation multimédia s'intéresse en premier lieu aux contenus visuels, typiquement des images décrivant des scènes naturelles et des vidéos provenant du monde des médias traditionnelles et des médias sociaux. Les contenus audio, notamment les contenus oraux provenant des médias et les bandes-son des vidéos,

sont également considérés bien que dans une moindre mesure. Les contenus textuels ont quant à eux été largement étudiés dans la communauté recherche d'information, et leur indexation n'est en générale pas considérée comme relevant de l'indexation multimédia. Le traitement de la multimodalité, *i.e.*, comment décrire et indexer des contenus combinant plusieurs modalités ou comment passer d'une modalité à l'autre (par exemple pour générer automatiquement des descriptifs textuels d'images [Vinyals et al., 2015–]), rend cependant ses barrières floues : ainsi l'indexation multimodale s'appuie largement sur des techniques textuelles d'indexation et de recherche d'information ; il en va de même pour l'indexation d'image dès lors que l'on est capable de décrire les images par des mots clés ou des étiquettes.

L'état de l'art proposé dans ce document se concentre en premier lieu sur les contenus visuels relevant du domaine du multimédia, et dresse un panorama des techniques permettant de les décrire et de les indexer de manière automatique. On intégrera également les aspects liés à la multimodalité permettant de faire le lien entre images et textes.

b. Présentation de la BnF □

La Bibliothèque nationale de France (BnF) est un établissement public à caractère administratif sous tutelle du ministère de la Culture. Selon les termes de son décret de création (n° 94-3 du 3 janvier 1994), la BnF a pour mission de

- de collecter, cataloguer, conserver et enrichir dans tous les champs de la connaissance, le patrimoine national dont elle a la garde, en particulier le patrimoine de langue française ou relatif à la civilisation française.
- d'assurer l'accès du plus grand nombre aux collections, sous réserve des secrets protégés par la loi, dans des conditions conformes à la législation sur la propriété intellectuelle et compatibles avec la conservation de ces collections.
- de préserver, gérer et mettre en valeur les immeubles dont elle est dotée.

Outre la constitution et la conservation des collections, la BnF doit les communiquer au public, tout en respectant les impératifs de ces premières missions, notamment ceux de conservation. Dans cette perspective, la BnF organise cette communication en sélectionnant le public par le biais de règles d'accréditation, mais aussi en ne communiquant parfois que la reproduction des documents les plus fragiles, de plus en plus nombreux à être numérisés et accessibles sur Gallica, sa bibliothèque numérique. En libre accès, elle regroupe des

livres numérisés, des cartulaires, des revues, des images, des enregistrements sonores, des cartes et une collection d'enluminures.

Au 17 août 2017, Gallica proposait à la consultation en ligne 4 252 443 documents, avec un rythme de 1 500 pages numérisées par jour, et 4 032 205 documents, dont 694 319 livres, 1 773 014 fascicules de presse et revues, 949 240 images, 80 084 manuscrits, 102 371 cartes, 44 160 partitions, 35 212 documents sonore, 353 769 objets et 18 vidéos.

(source:Wikipédia)□

La BnF désire améliorer le service et l'expérience utilisateur de sa plateforme numérique Gallica.

Pour répondre aux besoins d'accès à leur collection, la BnF souhaite mettre en place de nouveaux outils permettant d'enrichir et de valoriser ses contenus, notamment la mise en œuvre d'un outil de fouille d'images. Cet outil de recherche devra permettre aux utilisateurs d'explorer la collection en utilisant des indices de contenu (recherche par images similaires par exemple) ou des indices sémantiques (recherche par concepts).

c. Les corpus images de la BnF

Le corpus des images de la BnF est très varié. Ces corpus sont identifiés selon leur département d'origine et sont donc très hétérogènes. A titre d'exemple, nous présentons ici un panel du type d'images qui sont consultables et qui devront donc être valorisées via l'outil de fouille d'images. Ces corpus d'images, extraits de Gallica, ont été fournis par la BnF.

département des manuscrits

Le corpus des manuscrits est composé de 2 catégories :

- Les manuscrits autographes modernes sont des numérisations intégrales de manuscrits autographes d'auteurs de langue française, modernes et contemporains.
- Les iconographies figurant soit des animaux (mammifères), soit des activités humaines (chasse, construction, art militaire, tournoi, liturgie, (scène issue d'une) fable, fête)



réserve des livres rares

Ce corpus regroupe 3 sous-corpus :

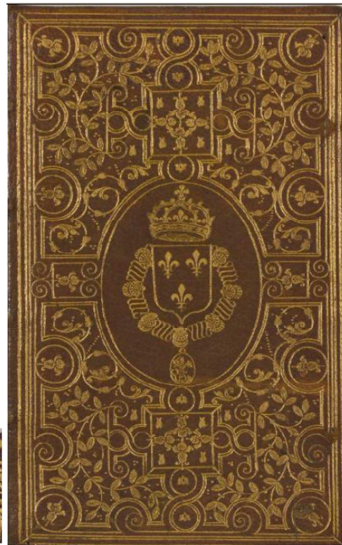
- L'iconographie des incunables issue de livres imprimés avant 1501. Les illustrations sont des gravures sur bois.



- Le matériel typographique (alphabet de lettrines, ornements...) sur les ouvrages imprimés entre 1527 et 1544.



- Les reliures à décor



Département des estampes

Ce corpus regroupe 2 sous-corpus :

- Les gravures :



- Les iconographies :
 - d'animaux
 - du second empire



DSC/NUM

Ce corpus est constitué d'un jeu d'illustrations extraits de la bibliothèque numérique Gallica avec pour thématique la Première guerre mondiale. Il couvre la période 1910-1920, plusieurs collections numériques (images, livres, presse) et divers genres (principalement gravure, dessin et photo). Pour les corpus imprimés (presse et livre), les illustrations sont déjà identifiées dans la page à partir des informations fournies par l'OCR et elles sont enrichies avec des descripteurs textuels (texte autour de l'illustration, légende de l'illustration).

Drawings (2024)



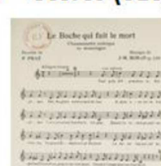
Photos (2449)



Advertisings (364)



Scores (616)



Comics (212)



Handwritings (64)

Mariam à la Coca
c'est bien mieux quinqu
à coté, la sainte Coca.
ROCHARD

Engravings (1133)



Maps (282)



Ornaments (35)



Covers (86)



Blanks (178)



Texts (378)

Les tsaristes, échoués que des vagues
normes s'abattent sur les navires,
après avoir tués Andrey et divers
autres navires, subirent de graves avari-
es qui furent balayés, un épais nu-
age de fumée.
On entendit des cris épouvantables,
un silence tragique se fit.
Au lever du soleil, le désastre apparut d
suite son horreur. La ville entière n'y
fut qu'un amas de débris d'où sui-
virent seulement les murs de l'Hôtel-de-V

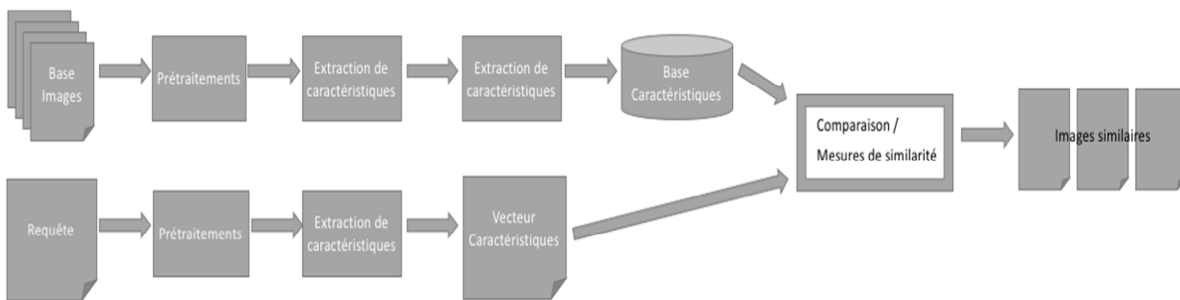
d. Architecture général d'un système de recherche

La qualité de consultation des bases d'images dépend directement de la précision et de la triviale avec lesquels un utilisateur va pouvoir trouver des images qu'il trouve intéressante. Il existe donc une multitude de méthodes permettant ces opérations. Selon la littérature, la recherche d'images est classifiable selon trois catégories :

1. recherche par navigation libre dans la base d'image : l'ensemble des images de la base sont présentées à l'utilisateur qui la parcourt pour rechercher une image. Cette technique n'est que peu utilisée puisqu'elle impose de naviguer longtemps au sein de la base sans garantir de résultat à l'utilisateur.
2. recherche par mots-clés : le principe général consiste à associer des mots-clés aux images (manuellement ou (semi-)automatiquement) et à rechercher des images à partir de ces mots-clés.

- recherche par le contenu : plus communément appelé Content-Based Image Retrieval (CBIR), cette technique utilise des méthodes qui décrivent le contenu des images.

Dans ce rapport, nous nous focaliserons sur les méthodes permettant une recherche d'images par le contenu. Comme illustré par la figure, un système d'indexation par le contenu consiste tout d'abord en une première phase, hors ligne, visant à extraire les vecteurs de caractéristiques des images de la base et les stocker. La deuxième, en ligne, permet à l'utilisateur de faire une recherche en présentant une image requête par exemple.



Au final, le système de CBIR retournera un ensemble d'images jugé pertinent.

Toutes ces techniques sont développées pour proposer quatre types de services aux utilisateurs :

- Navigation dans des catégories prédéfinies** : l'utilisateur navigue dans la base en choisissant des catégories pré-définies au moment de la construction de la base. Cette navigation impose de connaître les différentes catégories avant d'indexer la base et que toute la base soit indexée sur ces catégories.
- Recherche par l'exemple** (également connue sous l'appellation QBE ou Query by Example en anglais) : l'utilisateur propose une image en exemple de ce qu'il recherche. Après avoir résumé le contenu de l'image, le système recherche les

images les plus similaires, du point de vue du résumé (descripteurs de l'information contenue par les pixels).

- **Recherche par esquisse** : approche similaire à une recherche par l'exemple mais dans ce cas précis, l'utilisateur ne propose pas une image mais dessine une esquisse de ce qu'il recherche. L'esquisse peut-être basée sur une forme particulière ou une couleur comme le propose RetrievR (<http://labs.systemone.at/retrievr/>), service du site FlickrR (<http://www.flickr.com/>)
- **Spécification de caractéristiques visuelles** : l'utilisateur précise implicitement des caractéristiques qui résument l'image d'un point de vue bas-niveau. C'est-à-dire que les caractéristiques précisées n'ont pas de sens direct avec le contenu de l'image. Par exemple, l'utilisateur précise la taille, la couleur moyenne ou l'aspect des textures de l'image.

Dans la suite de ce document, nous présenterons les techniques d'analyse et d'indexation d'images. L'objectif dans cette étape est d'extraire et de caractériser la structure d'images graphiques pour la navigation dans des masses de données. Cela pose plusieurs problèmes théoriques difficiles qui sont :

- **Génération de la représentation** : Il s'agit d'extraire de l'image des signatures caractéristiques qui la résument. Il existe différentes stratégies qui conditionnent la qualité de la navigation dans ces masses de données et le temps de traitement. Ces différentes caractéristiques peuvent être de différents types (cf. section 3 et 4)
- **Comparaison des représentations** : une fois une représentation extraite des images, il s'agit de **mesurer la similarité entre deux images** à partir de leur représentation. Suivant la nature de la représentation, on se trouve confronté à un **dilemme qualité/rapidité**.
- **Problématique de la masse de données** : une partie des enjeux portent également sur la navigation dans de **grandes masses de documents**. En plus de la problématique de la représentation et de leur comparaison, il s'agit de mettre en place une structuration optimale de l'information pour éviter la comparaison exhaustive.

e. Synthèse

L'étude des corpus montre bien entendu une grande hétérogénéité des images disponibles. Les corpus disposent chacun de leur spécificités :

- Qualité des images disponibles : cette qualité dépend bien entendu de l'origine du document (son époque, sa vétusté, sa technique de réalisation) mais également à la numérisation du support. Un prétraitement peut alors s'avérer nécessaire pour améliorer la qualité du document ou filtrer certains bruits.
- un document peut présenter plusieurs contenus qui nécessitent d'être segmentés au préalable : recueils, iconographies multiple, lettrines... Dans ce cas là, une étape de segmentation est alors nécessaires pour permettre une isolation des contenus à indexer.
- Certains corpus sont indexés avec des métadonnées, qui sont dépendantes du corpus. Les métadonnées sont donc très diverses et dépendent fortement de la nature même du corpus. Les métadonnées peuvent être par exemple des informations de localisation, une sortie OCR (corrigé ou non), les identifiants ARK et/ou IFN... Ces métadonnées peuvent être incomplètes. Leur enrichissement peut alors permettre une meilleure indexation.
- La finalité des documents est également très variée. Certains corpus sont, par exemple, à destination d'une consultation par le grand public. La notion sémantique est alors très importante puisque l'utilisateur va vouloir chercher ces photos d'un même lieu sur plusieurs époques, ou alors des articles de journaux traitant d'un sujet précis. Les professionnels BnF (historiens, linguistes, ...) ou les chercheurs en général vont vouloir une recherche peut-être plus précise : toutes les lettrines d'un même imprimeur ou graveur par exemple. La mise en place d'une unique fouille d'images susceptible de répondre à toutes ces requêtes est alors complexe à mettre en place.

Afin de répondre au mieux à toutes ces problématiques, le document présente les divers verrous aux problématiques exposés ci-dessus. La section 2 présente un état de l'art des méthodes de prétraitements et de segmentation. La section 3 présente le panorama des méthodes d'indexation sémantique et d'enrichissement des métadonnées des images. La section 4 introduit les techniques d'indexation par le contenu visuel. La section 5 présente une liste des logiciels et des boîtes à outils disponibles actuellement et la section 6, une liste des principaux acteurs oeuvrant sur ces domaines en France et en Europe.

2. Pré-traitement et segmentation (Extraction de la structure non labellisée)

Comme indiqué ci-dessus, l'objectif de l'analyse de la structure du document (ou *Document Layout Analysis* en anglais), est de pouvoir séparer les régions du document numérisé qui correspondent à des zones particulières (illustration, graphique, texte, etc.) et de modéliser l'organisation spatiale et structurelle de ces différentes parties extraites. En général, pour atteindre cet objectif une analyse globale du document est faite afin de s'assurer que l'image contient les éléments selon une disposition la plus classique possible (sans distorsions, lignes horizontales et bon contraste des contenus). C'est sous le terme de pré-traitement que sont regroupées toutes les méthodes permettant de faciliter la segmentation du document numérisé en régions d'intérêt (ROI) où sont extraites des caractéristiques pour atteindre la reconnaissance ou la labellisation fiable de ces ROI. Dans [Shafait 2011], une étude de l'effet du bruit présent au niveau des bords des documents, sur les performances des méthodes de segmentation de documents numérisés est réalisée et montre que ce problème est encore un problème ouvert du fait de la variété des conditions de numérisation.

2.1 Pré-traitement

Parmi les nombreux prétraitements envisagés nous distinguerons deux grandes catégories de pré-traitements en fonction de leurs objectifs, d'une part ceux qui relèvent de l'aspect spatial, d'autre part ceux concernant la couleur, même si bien souvent les images restent en niveaux de gris.

- Parmi les problèmes de positionnement du document dans l'image, on doit considérer l'orientation et les distorsions du document de manière à corriger les erreurs qui persisteraient à la suite du protocole de numérisation. Cette catégorie de problèmes est très importante notamment dans le cas de

capture numérique via des tablettes, téléphones portables etc. Les méthodes font le plus souvent appel à la détection de lignes (lignes de textes ou bords du document), qui évaluent les transformations inverses à effectuer pour restituer un document non déformé [Brown 2007], [Liang 2008], [Zhang 2008], [Beusekom 2010], [Luo 2011], [Takezawa 2017]. Certains scanners professionnels intègrent de tels processus dans la machine. Dans le cas de la détection des bords du document, les méthodes sont bien évidemment liées à une bonne détection des bords réels du document au sein de l'image de document, tâche qui peut se compliquer lorsque l'éclairage introduit une illumination non constante au moment de la capture. Les méthodes s'appuient sur la détection de lignes ou sur une analyse de la couleur.

- Même si de plus en plus de traitements sont adaptés aux images en niveaux de gris, la plupart des méthodes supposent que l'image à traiter a préalablement été binarisée, en particulier quand l'intérêt du document à traiter réside dans son contenu textuel. Si cette étape de binarisation était souvent résolue au niveau de l'acquisition par le scanner, les chercheurs ont développé, depuis plusieurs dizaines d'années, de nombreuses méthodes plus ou moins sophistiquées pour améliorer le processus. Cette étape est généralement utilisée pour séparer le fond de l'image du document des informations écrites ou tracées au premier plan. Des approches globales, avec un seuil fixé par l'utilisateur ou estimé automatiquement [Otsu 1979] bien que rapides se montrent souvent inefficaces dans le cas d'images de documents avec un fond non uniforme ou dans le cas de documents anciens qui peuvent présenter des taches, des trous ou comporter des traces de l'écriture présente au verso de la page considérée. De plus, les images de documents peuvent également contenir des éléments d'information variés (texte, images, graphiques etc.) qui ne présentent pas un niveau de gris homogène, un seuil global n'est donc pas adapté pour séparer les informations du fond. Dans ce cas, des approches locales avec la détermination de seuils locaux, ont été développées permettant ainsi de traiter des images bruitées et de régler le problème de contraste de luminosité sur une même page. Un grand nombre sont développées à partir de la formule de Niblack [Niblack 1985] (sensible au bruit situé dans le fond), [Sauvola 2000], [Kim 2002], [Khun 2009]. Une difficulté de ces méthodes locales réside dans le choix des valeurs des paramètres des méthodes et

notamment la définition de la taille de la fenêtre au niveau de laquelle se fait l'étude, elle dépend de la hauteur des lignes de texte.

La tendance des méthodes actuelles est de combiner plusieurs informations qui peuvent être de même nature mais recueillies à des niveaux d'observation différents [Moghaddam 2012], ou de natures différentes via des étapes d'extraction de contours [Howe 2013], [Su 2013], de paramètres de texture [Nafchi 2014], [Sehad 2015], de paramètres d'estimation du fond et/ou de la forme [Gatos 2006], [Rivest-Hénault 2011], [Su 2013]. Ces méthodes combinées reposent sur des approches coarse-to-fine, variationnelles (modélisations par équations aux dérivées partielles) [Rivest-Hénault 2011], [Hadjadj 2014] ou de minimisation énergétique [Howe 2013], probabilistes [Gatos 2006], [Rivest-Hénault 2011], [Gaceb 2013], [Su 2013], [Chan 2017], des approches issues de l'intelligence artificielle avec les systèmes multi-experts [Moghaddam 2015] ou des systèmes d'apprentissage qui permettent de déterminer automatiquement les valeurs optimales des paramètres des méthodes de binarisation [Su 2010], [Su 2012]. Enfin plus récemment, des approches de type Deep Learning ont fait leur apparition [Tensmeyer 2017], fondées sur un apprentissage des caractéristiques du fond des documents, et se montrent très compétitives.

Les résultats obtenus par ces méthodes combinant plusieurs sources d'information sont très bons comme en témoigne leur classement dans diverses compétitions.

Dans le cadre de documents couleurs, la plupart des méthodes passent par une conversion de l'image couleur en niveaux de gris avant d'opérer l'étape de binarisation. Une étude de l'impact de cette transformation couleur-niveaux de gris a été réalisée dans [Hedjam 2016]. Cependant, certaines méthodes travaillent dans l'espace couleur directement comme les approches locales fondées sur la détection de couleurs dominantes en appliquant un principe qui permet plus généralement de segmenter l'image selon les couleurs qui apparaissent [Verleysen 2013]. D'autres méthodes sont fondées sur du clustering adaptatif dans l'espace des couleurs [Leydier 2004], ce qui permet de prendre en compte les variations d'illumination et les problèmes de transparence. Elles nécessitent néanmoins une initialisation interactive pour déterminer le nombre de classes couleurs et à l'intérieur de chaque classe couleur quelques exemples de couleur. Par ailleurs dans un document où la couleur est utilisée, cette couleur apporte une information qui permet d'extraire des éléments comme les tableaux à

lignes ou colonnes matérialisées par des modifications dans la couleur du fond [Alheritiere16].

Le problème de la binarisation, même si d'importantes améliorations ont été obtenues, est un problème encore ouvert comme le montrent les compétitions organisées dans le domaine [Gatos 2009], [Gatos 2017], car malgré les efforts fournis aucune méthode ne permet de traiter de manière fiable tous les types de documents ni tous les types de dégradation.

2.2 Segmentation ou extraction du layout physique

Le layout physique d'un document correspond à l'organisation spatiale et structurelle des différentes parties constituant celui-ci. Il permet d'identifier partiellement le contenu du document sans en avoir analysé précisément les différentes parties. Il peut être complété par un layout logique (par exemple informations de titre, de résumé, de pied de page, de section, légende, liens explicites et implicites qui existent entre les parties, etc.) [Faure 2007], qui correspond à l'interprétation de ce qu'un humain a apporté à ce document. La multitude des documents possédant des contenus très hétérogènes en contenu et en disposition, constitue la principale difficulté de la détection et de l'analyse des layouts. Les approches présentées dans [Nagy 2000] [Trupin 2005] et [Journet 2008] montrent cette diversité des structures avec des tentatives de classification, respectivement pour les documents imprimés et documents anciens.

L'analyse du layout d'un document (qu'il soit logique ou physique) est importante dans la dématérialisation de ce dernier. En effet, connaître la nature des différentes parties d'un document permet de faciliter son indexation en prévision de sa recherche à l'intérieur d'une base de documents, mais aussi de faciliter sa visualisation et sa lecture sur un nouveau support comme un téléphone mobile ou une tablette. L'analyse du layout du document (*DLA*) peut également servir à la classification d'images de documents selon un critère d'apparence sans avoir d'*a priori* sur le contenu textuel de la base d'images à traiter, par exemple une page ne comportant que des images. Cette tâche de *DLA* fait régulièrement l'objet de compétitions où les images de test présentent des layouts de plus en plus complexes au fil des années [Antanacopoulos 2013], [Antanacopoulos 2015], [Antanacopoulos 2017].

Étant donné la variété et l'hétérogénéité des documents à traiter, nous excluons de cet état de l'art les méthodes à base de grammaires [Nagy 1984a], [Belaïd 1997] permettant de localiser les différents éléments d'un document, souvent développées pour extraire le layout logique mais qui présentent le principal inconvénient de nécessiter une certaine

connaissance des documents traités pour obtenir des résultats performants. Nous éliminerons également les méthodes nécessitant une intervention de l'utilisateur lors de la chaîne de traitement [Ramel 2006].

Dans la présente section, seul le layout physique sera abordé. L'analyse de ce layout physique permettra de classer les informations contenues dans les documents en vue de leur indexation, en fournissant notamment une labellisation (étiquetage sémantique) relative au type de media des ROIs extraites.

Quelques articles de revues de méthodes d'analyse de structure du document [Cattoni 1998], [Mao 2003] ou plus généralement sur la classification d'images de documents [CHE08] existent mais commencent à être relativement anciens. Ces études présentent néanmoins certaines méthodes qui ont servi de socle au développement d'améliorations ces dernières années. Nous allons donc rappeler les méthodes de base et montrer leurs évolutions plus récentes. D'un point de vue global, il existe des méthodes d'analyse du layout physique fondées sur la **classification** (selon des critères d'homogénéité liés à la couleur, à la forme par exemple), et des méthodes faisant intervenir une **segmentation** d'images suivies d'une étape de **labellisation**. Enfin, étant donné le nombre limité et connu de médias, des approches d'extraction par couches peuvent être mises en concurrence de manière à obtenir une segmentation finale consensuelle.

Pour les approches **sans segmentation** [Baird 2007] [Vieux 2012], [Cote 2014], fondées sur la classification, cette classification se fait généralement à partir d'un apprentissage. Plusieurs types d'apprentissage sont possibles :

- un apprentissage non-supervisé repose sur un choix pertinent de caractéristiques fondées sur la couleur [Baird 2007] dans les différents espaces possibles (RGB, HSL, etc), la forme [Vieux 2012], un mélange de caractéristiques de différentes natures [Cote 2014], etc. Ces caractéristiques sont choisies pour permettre de séparer au mieux les classes d'intérêt recherchées ;
- un apprentissage supervisé nécessite d'avoir un jeu de données labellisées pour l'apprentissage ;
- un apprentissage semi-supervisé dans lequel on dispose d'un jeu de données d'apprentissage seulement en partie étiqueté.

Ces méthodes présentent l'avantage majeur de ne pas nécessiter d'*a priori* sur les zones recherchées, mais en revanche les résultats fournis sont en règle générale moins précis que les méthodes associées à une phase de segmentation au niveau des contours des éléments du layout. En effet la classification se fait soit au niveau des pixels, soit au niveau de petites zones. Dans le premier cas ce sont les caractéristiques associées aux pixels qui sont classifiées et il apparaît dans les classes trouvées de

nombreuses composantes connexes ayant un nombre de pixels non significatif pour une ROI. Dans le cas de la classification de zones rectangulaires, rien ne garantit que la frontière de ces zones corresponde avec la frontière des ROIs recherchées. De plus, dans le cas de méthodes supervisées ou semi-supervisées, ces méthodes nécessitent un grand nombre d'images ou imagerie variées pour la phase d'apprentissage. Ces méthodes ont l'avantage de pouvoir s'appliquer sur des images en niveaux de gris ou même en couleur.

Dans les approches par couches, on cherche à extraire des éléments précis dans le contenu du document ou média, qui constitue ici une couche de l'image. Chaque média est extrait sans étape de segmentation préalable. Cette analyse par couche peut avoir lieu à trois niveaux : la détection (présence ou absence du média), la localisation (localisation précise du média), la reconnaissance (contenu de la couche reconstitué et labellisé). Ces approches sont par exemple utilisées dans des méthodes de séparation de textes imprimés/manuscrits [Grzejszczak 2012], [Hamrouni 2014], dans la détection et l'extraction des tableaux avec séparateurs matérialisés [Cesarini 2002] ou non [Shafait 2010], en noir et blanc ou à alternance de couleurs [Alheritiere 2016], la détection et l'extraction de logos [Jain 2012], [Le 2012]. Le problème de ces approches est que les couches ne sont pas toujours définies précisément notamment en ce qui concerne le niveau d'observation. Par ailleurs, les méthodes par couches performantes sont généralement spécialisées dans un type de documents, et s'appuient sur des règles de mise en page de ces documents. Leur avantage principal est que l'on peut extraire les informations à plusieurs niveaux, texte, image, graphique, tableaux etc., et que ces informations obtenues de manière indépendante peuvent être confrontées de manière à trouver un consensus et reconstruire un résultat global, au prix évidemment d'un temps de traitement plus important.

Dans le cadre des approches fondées sur une étape de **segmentation**, cette dernière consiste le plus souvent à séparer les zones textuelles des zones graphiques du document. Les zones textuelles participent à la détection des mots, lignes, colonnes, paragraphes, alors que les zones graphiques permettront de séparer les symboles, logos, signatures, graphiques, lignes séparatrices etc.

Dans le domaine de la segmentation, il n'existe pas de méthode universelle, mais plutôt des familles de méthodes adaptées à des catégories d'images. Les images de documents n'échappent pas à cette règle. Concrètement les développements actuels posent la question du choix de l'échelle d'observation (pixel, groupements de pixels

répondant à des critères de voisinage, d'organisation spatiale, de niveaux de gris etc.) et des caractéristiques. Certaines méthodes font le choix d'une multiplicité de caractéristiques, quitte à tomber dans la malédiction de la dimension, d'autres préfèrent choisir un nombre restreint de caractéristiques bien adaptées au problème à régler.

Parmi les méthodes de segmentation d'images suivies d'une labellisation, on rencontre trois grands types d'approches :

- Les méthodes descendantes (top-down) :

Ces méthodes font appel à une décomposition du document initial en zones plus petites reposant sur la satisfaction d'un critère (XY-cut [Nagy 1984b] et ses variantes [Meunier 2005], [Coppi 2014]). Elles s'appliquent aussi bien sur des images binaires que sur des images en niveaux de gris, et s'avèrent très efficaces dans le cas de *layouts* simples avec des blocs rectangulaires, mais sont sensibles à l'inclinaison du document. Cette gestion de l'inclinaison peut avoir été traitée au niveau de la phase de prétraitement. En revanche, elles ne peuvent traiter des documents avec layout complexe.

Les blocs obtenus peuvent ensuite être classifiés par une machine à vecteur support (Support Vector Machine ou SVM) [Coppi 2014].

- Les méthodes ascendantes (bottom-up) :

Ces méthodes se fondent sur des agrégations successives de pixels avec par exemple la méthode RLSA (Run Length Smoothing Algorithm, [Wahl 1982], [Grzejszczak 2012], [Hamrouni 2014]), particulièrement adaptée à la segmentation de zones textuelles. Le RLSA ne s'applique que sur des images binaires de documents et donc une étape de binarisation est nécessaire pour traiter les images de documents en niveaux de gris. Elle consiste à « noircir » les espaces blancs entre deux pixels noirs (appartenant généralement à des composantes de la forme, alors que les pixels blancs correspondent à des éléments du fond), dont la distance est inférieure à un seuil. La valeur de ce seuil fixé empiriquement permet d'agréger des composantes textuelles à plusieurs niveaux (lettre, mot, ligne, etc.) mais permet aussi d'agréger des pixels appartenant à des illustrations (images ou graphiques). Pour pouvoir traiter des éléments différents ou à des niveaux d'observation différents, il faudrait que cette valeur de seuil soit adaptative. Par ailleurs elle est sensible à l'inclinaison du contenu du document.

Un deuxième type d'approches ascendantes est fondé sur le diagramme de Voronoï, qui permet un découpage du plan à partir de certains points particuliers ou germes, où la fusion des cellules peut être fondée sur une étude statistique des distances entre les

composantes textuelles par exemple [Kise 1998], [Winder 2011]. A partir des frontières des composantes connexes, un premier diagramme de Voronoï est calculé. Ensuite la fusion des cellules de ce premier pavage est fondée sur une étude statistique des distances entre les composante (espaces inter-lettres, inter-mots, inter-lignes). Son avantage majeur est la possibilité de traiter des documents même avec une forte inclinaison ou présentant un layout complexe.

Le troisième type de méthodes dans cette catégorie des approches ascendantes, regroupe les méthodes fondées sur la multi-résolution où des opérateurs différents sont appliqués en fonction d'une échelle d'observation qui peut varier de grossière à fine ([Bloomberg 1991] travail à l'origine de la librairie open source Leptonica, méthodes fondées sur l'analyse des zones « blanches » et des zones « occupées » qui représentent une certaine dualité (analyse de zones blanches sous forme d'espaces polygonaux pour autoriser le traitement de documents inclinés [Antanacopoulos 1998], méthodes fondées sur l'analyse de composantes connexes selon un ou plusieurs critères (alignement, espacement, etc.), [Vauthier 2012]. Ces méthodes nécessitent également une étape de binarisation préalable du document image, qui les rend souvent sensibles au bruit et dépendantes du résultat de cette étape.

- Les méthodes hybrides :

Ces méthodes combinent les approches descendantes et ascendantes. Nous pouvons citer ici la méthode des Tab-stops [Smith 2009], Les Tab-stops sont les composantes qui commencent ou terminent les lignes de texte. Les trouver permet, en calculant leur alignement, de trouver les lignes et ainsi de segmenter un document. La première étape de cette méthode consiste à 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. Une analyse des composantes connexes permet alors de détecter des texte-candidats puis de déterminer parmi ceux-ci les Tab-stops et plus particulièrement ceux qui peuvent être appariés en paires formant le début et fin de chaque ligne. Enfin selon des critères de taille et d'espacement inter-lignes, les lignes détectées pourront être regroupées en blocs. Cette méthode à l'origine de la librairie Tesseract, donne des résultats intéressants mais est fortement perturbée en présence de tableaux dans les documents.

Une approche similaire est utilisée dans AOSM [Ha 2016]), où une première subdivision de l'image est ensuite remise en cause par des fusions des ROIs, initialement obtenues,

en fonction de certaines règles ou critères à respecter au niveau des layouts (présence de zones de fond, de séparateurs, etc.).

Enfin, nous pouvons noter parmi ces méthodes une modélisation de l'image d'une part, et de son fond par des segments de droites de direction quelconque. Le regroupement de ces segments en fonction de leur couleur, de leur longueur ou de leur orientation permet de reconstruire les ROIs textuelles, les séparateurs, tableaux, les images, les graphiques [Alheritiere17].

Ces méthodes, du fait de la prise en compte de plusieurs sources d'informations, et notamment de l'exploitation de la dualité fond/forme permettent de mieux s'affranchir de la variabilité observée au niveau des images de documents.

Enfin plus récemment également, des approches fondées sur un apprentissage par des réseaux de neurones simples [Chen 2017] ou profonds (Deep Learning) ont fait leur apparition [XU17], notamment pour la segmentation de pages de documents anciens (texte, décoration, commentaires). Ces réseaux permettent d'apprendre à partir d'informations brutes en grandes quantités, en réalisant 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. Elles ne sont pour l'instant pas en tête des compétitions de segmentation de pages de documents, où les systèmes à base de règles de décision ou de combinaison de plusieurs sources d'information semblent encore pour l'instant plus performants [Antanacopoulos 2017].

3. Indexation sémantique des images

L'indexation sémantique des images consiste à détecter automatiquement des concepts dits sémantiques, i.e., facile à comprendre et à interpréter pour les utilisateurs, dans les images afin de créer un index sémantique à partir des concepts détectés. Au cours des dernières années, avec l'explosion des architectures de Deep Learning, ce domaine de recherche a été révolutionné et a atteint des performances exceptionnelles dans des tâches difficiles [Goodfellow2016].

Dans cette partie du document, nous examinerons les dernières technologies et applications qui ont été proposées et qui ont été rapidement acceptées par la communauté de Vision par Ordinateur, ce qui, selon nous, aurait du sens dans le contexte d'une bibliothèque numérique qui souhaiterait améliorer l'expérience de l'utilisateur lorsqu'il cherche un certain contenu.

On visera en premier lieu l'application sur des images de documents, et ensuite quelques approches sur l'analyse de photographies.

L'étape de labellisation/étiquetage a pour objectif d'affecter un label/concept (photo, portrait, carte, illustration, etc.) sémantique (*i.e.*, plus aisé à comprendre et à interpréter pour les utilisateurs) aux différentes zones constituant le contenu du document. De telles zones résultent de l'extraction du layout physique. On considérera ici que l'approche employée pendant la segmentation a déjà proposé une première labellisation « grossière » des zones du document afin de séparer les zones de texte, des images, etc. Nous nous intéressons ici dans cette étape de labellisation à un raffinement de ces labels : une zone identifiée comme « image » par la segmentation peut alors, dans cette nouvelle étape, être étiquetée comme « photo », « portrait » ou « carte », etc. De tels labels offrent une nouvelle représentation du contenu du document avec un plus haut niveau de sémantique et peuvent être employés pour enrichir les métadonnées d'un document lors des phases ultérieures d'indexation et de recherche par le contenu.

On trouve dans la littérature plusieurs familles d'approches pour la labellisation de zones de documents. Elles reposent généralement sur des approches de classification supervisée issues de l'apprentissage automatique. La méthode de classification apprend alors un

modèle sur un jeu de données dont on connaît déjà la classification pour produire un modèle de prédictions. Les approches de labellisation sont de natures différentes suivant qu'elles cherchent à labelliser directement des pixels du document ou des régions issues de l'étape de segmentation. Quelle que soit la primitive considérée (pixel ou région), un point commun aux approches de labellisation est de proposer un label à affecter en fonction des caractéristiques extraites (cad, SIFT, SURF, GIST) du contenu de l'image à partir de la primitive choisie. Les approches basées régions permettent d'inclure des caractéristiques de plus haut-niveau (comme des informations géométriques relatives aux régions à étiqueter) qui peuvent s'avérer plus discriminantes suivant les labels considérés.

Des méthodes spécifiques ont également été développées pour certains types de labels (portraits, cartes, etc.). Par exemple pour la labellisation de « portraits », les méthodes peuvent prendre en compte des caractéristiques spécifiques des visages comme le rapport largeur/hauteur, la présence des yeux et de la bouche dans le cas des visages, etc. D'autres approches permettent un raffinement des labels considérés via des mécanismes semi-interactifs d'apprentissage comme le retour de pertinence d'un utilisateur ou l'apprentissage incrémental. Aujourd'hui, l'avènement des réseaux de neurones avec des architectures de plus en plus complexes (AlexNet, VGG, GoogleNet, RCNN, Fast RCNN, Inception) offre de nouvelles perspectives pour la prise en compte de nombreux labels avec un niveau de sémantique de plus en plus spécialisé.

Un problème sous-jacent aux approches de labellisation sémantique est la très grande similarité visuelle entre certains concepts représentés dans les images (comme par exemple différencier une zone du document comportant une « photo » d'un « portrait », etc.). Des approches spécifiques pour la gestion de ce problème ont été proposées dans la littérature (on parle en anglais de Fine-grained Recognition). Une stratégie possible consiste à entraîner un classificateur à prédire pour des zones du document des labels sémantiques appartenant à un vocabulaire contrôlé et organisés hiérarchiquement (relation d'hyponymie, hyponymie). De telles relations sont ensuite utilisées pour minimiser l'erreur commise par le système : si le classificateur hésite entre le label « zèbre » et « vache » il prédira le label « mammifère » qui est moins spécialisé dans la hiérarchie de concepts. Un autre problème inhérent à l'utilisation d'approche supervisée pour la prédiction d'un très grand nombre de concepts est la difficulté de produire des bases d'images annotées, nécessaires pour l'entraînement des modèles. Les concepts sémantiques sont généralement donnés par un expert pour étiqueter l'image de document de manière globale (ce document contient une zone représentant une carte) et non directement pour la zone du document concernée. Une tendance actuelle consiste à considérer un tel problème comme

un problème de classification multi-labels et des résultats prometteurs ont déjà été obtenus pour la classification d'images naturelles.

Les labels considérés peuvent être variés et ces derniers peuvent ou non être inclus dans un vocabulaire contrôlé (ontologie, etc.) permettant par la suite de raisonner (inférence, etc.) et de construire un index sémantique à partir des concepts détectés. Ce type d'approche a déjà été utilisé avec succès pour la labellisation d'images naturelles de scènes : la base ImageNet contient un grand nombre d'images annotées avec 80 000 concepts définis à partir de l'ontologie WordNet. Ces nouvelles approches ont permis des progrès spectaculaires dans le domaine de la labellisation automatique et pourraient être adaptées pour le problème spécifique de l'analyse des documents de la BNF.

Le problème d'étiqueter des images à partir de l'addition de métadonnées (descripteurs thématiques) aux images, afin de cataloguer des photographies, pourrait être automatisé. Au cours des dernières années, des modèles de vision par ordinateur pour la reconnaissance de contenu dans des images ont été proposés pour être appliqués dans ce scénario afin d'automatiser le processus de catalogage thématique des images. Plus précisément, l'utilisation de modèles de Deep Learning a produit des algorithmes capables de surpasser la performance humaine dans la tâche de base de la classification des images [He2015] - une étape inimaginable il y a seulement cinq ans - et a ouvert la voie à des tâches encore plus complexes.

Comme base de test, la base de données ImageNet comporte un grand nombre d'images indexées en fonction de 80000 catégories définies à partir des synset de l'ontologie WordNet. La notion de concept est évidemment à définir en fonction des applications visées. On dispose cependant aujourd'hui de données et de technologies pour un grand nombre de concepts présents dans des scènes naturelles. Des sous-ensemble des catégories ImageNet ont été largement étudiés dans le cadre de campagnes d'évaluation telles ImageNet Large Scale Visual Recognition Challenge (ILVCSR) portant sur quelques milliers de concepts (Russakovsky et al., 2015) ou TRECVID Semantic Indexing visant la détection d'entre 500 et 1 000 concepts selon les années [Awad et al., 2016]. Ces campagnes d'évaluation ont permis des progrès spectaculaires dans le domaine et la mise à disposition de logiciels ou toolboxes permettant la détection de nombreux concepts dans les images de type scènes naturelles. Les technologies de détection et de reconnaissance de visage permettent également d'indexer des contenus selon les personnes présentes dans une image. On pourra également ranger dans cette catégorie les techniques de détection et de reconnaissance de logos dans les images.



Figure 3.A: Exemple de l'étiquetage automatique sémantique sur la base de ImageNet

Des approches similaires ont été proposées sur d'autres types d'images, notamment, la base de données *Places*, du M.I.T. qui vise à représenter des lieux physique a été utilisée pour de manière automatique reconnaître quelle scène représente une image [Zhou2014], où même proposer un système de recherche d'images par similitude sémantique [Gordo2016], où là, on ne vise pas a simplement étiqueter de manière automatique les contenus d'une image, mais de lister de manière ordonnée les images de plus à moins similaires étant donnée une image requête.

Si les méthodes antérieures visent à donner une liste d'étiquettes sémantiques qui décrivent les contenus des images, de nouvelles approches qui reconnaissent des objets sémantiques et qui sont capables de les localiser dans les images on été aussi proposés. Des méthodes comme YOLO [Redmon2016] ou SSD [Liu2016] permettent de détecter où dans l'image apparaissent les objets classifiés.

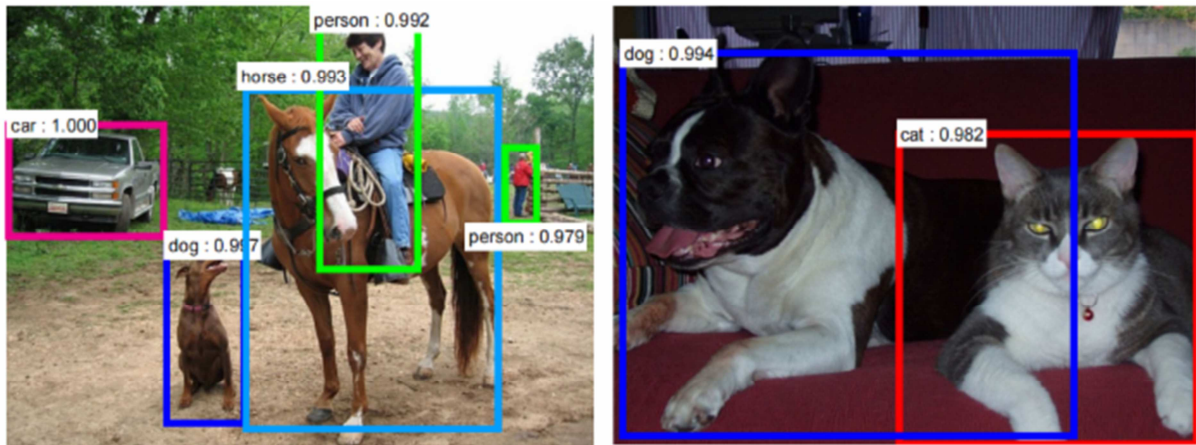


Figure 3.B: Exemple de classification et localisation d'objets sur la base ImageNet

Toutes les méthodes antérieures ont besoin d'avoir une grande quantité de données étiquetées pour pouvoir apprendre, et donc, fonctionner de manière automatique après. Or sur certains scénarios, avoir cette grande quantité d'images étiquetées peut ne pas être viable. Les recherches récentes développées au CVC mettent en avant l'idée d'intégrer des images et du texte dans un espace commun en exploitant une collection à grande échelle de documents multimodaux (texte et image). Ce cadre d'intégration peut être utilisé pour effectuer différentes tâches, telles que l'apprentissage autodirigé des caractéristiques visuelles et la récupération d'images multimodales [Gomez2017], ou même pour générer des lexiques contextualisés pour la reconnaissance de texte de scène [Patel2016]. Pour les objectifs de ce projet, le même cadre peut être utilisé pour établir des corrélations contextuelles sémantiques entre le matériel existant dans les collections d'images et les documents d'autres sources, tels que des requêtes textuelles libres, des images participatives ou même un corpus textuel entier.

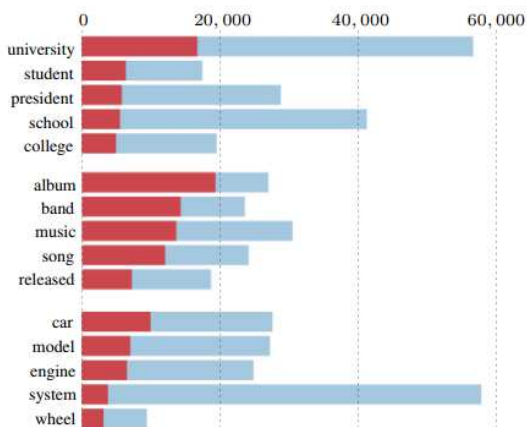


Figure 3.C. Sujets Sémantiques extraits de documents textuels et associés automatiquement à des images.

Une autre famille d'approches qui pourraient être intéressantes dans le domaine d'une bibliothèque numérique pourrait être celle du sous titrage automatique d'images. Le problème du sous-titrage automatique des images va plus loin que la simple reconnaissance du contenu individuel apparaissant dans les documents. Une fois les éléments visuels des images reconnus, les méthodes proposées sont capables d'offrir une description en langage naturel de ces contenus. Par conséquent, générer automatiquement des sous-titres dans les images. Dans la figure 3.D, nous montrons les résultats préliminaires obtenus par le CVC ces derniers mois en adaptant la méthode proposée dans [Vinyals2015] à l'application spécifique des photographies historiques dans les archives numériques. Ce problème est très intéressant et ouvre une série de lignes de recherche à prospecter.



A group of people standing around a white table	A store front with a bunch of signs on it
---	---

Figure 3.D. Sous-titrage automatique d'images

[OBJ]

D'une part, les résultats ont été mis en œuvre en utilisant des bases de données d'images récentes et de vieilles images des collections d'Europeana comme base de formation pour l'apprentissage. Pour le moment, le système est en mesure d'offrir des descriptions en anglais. Afin qu'il fournisse des sous-titres dans d'autres langues comme le français, un bon nombre d'éléments doivent être étiquetés dans cette langue ou voir comment le système peut être adapté de sorte "qu'il puisse parler d'autres langues".

Un autre défi correspond au type de descriptions générées. À l'heure actuelle, les systèmes de l'état de l'art sont capables de décrire uniquement le contenu visuel des images. Pour que ces systèmes fournissent beaucoup plus d'informations, il faudrait étudier comment on peut faire afin d'inclure des informations provenant d'un contexte historique et d'entités spécifiques.

Finalement, très récemment, la communauté de vision par ordinateur s'est intéressé au problème de répondre à des questions (plutôt simples), de manière automatique où la réponse se trouve dans une image. Ce problème est intéressant puisque le modèle doit en premier lieu comprendre qu'est-ce que l'utilisateur est en train de demander en langage naturel, puis, analyser l'image pour trouver la bonne réponse pour finalement la produire aussi en langage naturel. Les derniers travaux [Goyal2017] donnent déjà des résultats assez intéressants sur des cas d'usage simples.

Who is wearing glasses?
man
woman



Where is the child sitting?
fridge
arms



Is the umbrella upside down?
yes
no



How many children are in the bed?
2
1



Figure 3.E: Exemple d'images utilisées pour les systèmes de réponses visuelles automatiques

4. Indexation par le contenu visuel

4.1 Introduction

À l'inverse de l'indexation sémantique, l'indexation du contenu vise à s'affranchir d'une description sémantique du contenu de l'image pour ne s'intéresser qu'à la similarité visuelle entre deux images ou entre deux portions d'images. La notion de similarité visuelle peut bien évidemment porter sur différents aspects de l'image, de la couleur aux objets présents.

Nous commençons tout d'abord par un état des lieux des approches utilisées pour générer la représentation des images.

4.2 Caractéristiques

La recherche d'objets dans des images est très largement utilisée par les particuliers et les industriels. Les premiers systèmes mis au point reposaient exclusivement sur la recherche de motifs prédéfinis à partir de corrélations [Smeulders 00]. Ces approches présentaient l'inconvénient de ne pas être invariantes aux changements d'illuminations, d'échelles ou aux rotations.

Afin d'apporter une réponse à ce problème, des travaux de recherche ont proposé d'extraire des caractéristiques globalement sur l'image [Yang 08]. Ces descripteurs résument le contenu des images par des statistiques calculées sur les valeurs des pixels, et présentent l'avantage d'être invariants en rotation et changements d'échelles quand ils sont normalisés. La recherche consiste alors à une comparaison de ces descripteurs. De nombreuses caractéristiques ont été proposées dans la littérature mais celles-ci ne permettaient pas de décrire l'organisation spatiale des objets contenus dans les images.

Une alternative aux recherches d'images décrites globalement consiste à extraire des éléments structurants des images et à décrire leur agencement. De nombreuses approches ont été proposées en utilisant des segments, des contours, ou des régions comme éléments structurants. La description revient alors à indiquer les relations spatiales ou topologiques qui unissent ces éléments. Les caractéristiques spatiales fonctionnent bien pour certaines catégories d'images (telles que les images de symboles, les images composées de régions identiques). Dans le cas général d'images de scènes naturelles, ou composées de plusieurs

éléments différents, il devient difficile de pouvoir désigner les primitives structurelles, de les segmenter facilement et de décrire leur agencement.

Pour répondre à ce problème, de nombreux travaux, basés sur des des approches semi-locales, ont vu le jour. Une approche semi-locale va tout d'abord extraire des zones d'intérêts et les décrire (en utilisant des descripteurs semblables à ceux des approches globales). C'est la concaténation de ces descriptions qui permet de décrire une image. Dans certains cas particuliers, la description de l'agencement des zones d'intérêts est conjuguée à la description des zones elles-mêmes.

Nous présentons un panorama des trois grandes catégories d'approches :

Approches par description globale

Très largement utilisées dans la littérature, ce type de méthode vise à calculer des attributs sur l'ensemble des pixels de l'image pour les décrire. Elles décrivent les images en résumant son contenu à partir de la couleur [Yue 11, Wang 11], de corrélations [Shechtman 07], de traitements particuliers (redimensionnement d'images [Torralba 08], caractérisation des sous-éléments d'objets [Coustaty 08]), de caractéristiques de formes génériques [Tabbone 06, Zhang 02], ou d'approches basées sur la texture [Lowe 99, Choksuriwong 05, Journet 08]

Cette description est généralement représentée sous forme d'un vecteur de caractéristiques qui tente de décrire de la manière la plus unique et la moins ambiguë possible chaque classe d'images. Ce vecteur est alors utilisé pour indexer l'image, ou pour comparer des images via des distances entre vecteurs. Elles ont longtemps été les plus performantes dans des processus de reconnaissance de formes hors contexte.

Approches par description spatiale

Bien qu'elles permettent de résumer globalement les images, les approches globales ne permettent pas d'identifier la position de l'information (perte des repères spatiaux). Afin de pallier ce problème, de nouvelles signatures, souvent connues sous le nom de signatures structurelles ou topologiques, proposent des réponses à ce problème. Le principe de toutes ces méthodes repose sur trois étapes fondamentales :

Extraction de primitives

Ces primitives peuvent être des pixels, des segments, des zones homogènes, des régions, ... ([Uttama 05]). L'extraction de primitives a été appliquée sur des images naturelles (segmentations en régions homogènes), mais également sur des images de documents anciens. Dans ce cas, les objets présentent une structure forte, structure inhérente à leur construction. Par exemple, dans [Coustaty 08], le but est de comparer des symboles qui sont naturellement très structurés (composés de segments). Les deux approches proposent soit d'extraire les segments avec une transformée de Hough adaptée, soit d'extraire les polygones formés par les segments. Cependant, d'autres approches ont proposé de décrire l'organisation spatiale de zones particulières d'une image (position du ciel par rapport à la mer, position des yeux dans un visage, ...). Les primitives sont alors extraites soit suite à une division de l'image en sous-blocs de taille fixe (quadrillage appliqué sur l'image), soit selon des critères d'homogénéité (principe identique à celui utilisé dans les algorithmes de sélection de régions uniformes ou texturées vu précédemment).

Codage de l'agencement entre ces primitives

L'idée générale des signatures structurelles et spatiales est de décrire l'agencement et la position des formes qui composent un objet et/ou une image au sens large. Une fois les zones d'intérêt extraites, plusieurs solutions ont été proposées pour coder les relations qui existent entre ces différentes primitives élémentaires. Le codage des relations spatiales a fait l'objet de nombreuses études, en particulier dans le cadre des informations géo-référencées, mais également en indexation et reconnaissance des formes. Les approches proposées cherchent à caractériser la position spatiale d'une région avec une autre. Ainsi, elles permettent de connaître la position relative d'une région par rapport à une autre (au-dessus, à gauche, ...), ou si les régions possèdent des zones en commun (superposition, inclusion, ...). Des approches se sont servi de la topologie implicite des graphes pour représenter les formes qui composent une image. Ainsi, les graphes d'adjacence des régions [Bodic 09] décrivent l'agencement des régions qui composent une image (les noeuds correspondent aux régions et les arcs aux liens entre les régions). Enfin, des approches plus adaptées aux documents [Mas 10] ont proposé d'analyser la structure de primitives qui composent un symbole ou un objet. Cette structure est représentée à l'aide de grammaires 2D, ou d'identifiants propres à la relation qui unit deux éléments structurants (identification du lien existant entre deux polygones ou segments par des relations).

Représentation du codage

Enfin, une fois la liste des relations calculées sur l'ensemble des primitives élémentaires, toutes ces relations sont représentées sous forme d'arbres, de grammaires [Mas 06] ou de graphes [Bunke 11]. L'utilisation de structures permet généralement de retrouver intuitivement la forme de départ, et facilite la comparaison entre deux images. Ainsi, la reconnaissance ou l'indexation se fait par comparaison de ces structures (similarités de structure, isomorphismes entre graphes, ou dérivations de grammaire).

Approches par description locale

Décrire une image globalement présente l'avantage de conserver un maximum d'information sur son contexte mais ne permet de description en détails. Pour pallier ce problème, plusieurs approches ont été développées pour identifier des points d'intérêt dans une image (zones visuellement importantes) et décrire leurs alentours. C'est le but des approches locales (points d'intérêt) ou semi-locales (zones d'intérêt). L'idée revient à utiliser les approches globales de manière locale sur une sous-partie de l'image.

Leur bon fonctionnement repose bien évidemment sur l'extraction et la sélection de ces sous-parties pour qu'elles soient les plus pertinentes possible dans la description des images. Ces zones peuvent, en général, être caractérisées par :

- une définition claire, avec des fondements mathématiques de préférence
- une position bien déterminée dans l'espace image ;
- la structure locale de l'image autour du point d'intérêt et riche en terme d'information sur son contenu, de telle sorte que l'utilisation de ce point réduit la complexité des traitements suivants dans le système de vision ;
- une stabilité aux perturbations locales et globales de l'image, incluant les déformations topologiques et de perspectives (transformations affines, changements d'échelles, rotations, translations), comme les changements d'illuminations et de contrastes, de telle sorte que les points d'intérêt possèdent une forte reproductibilité ;
- une intégration optionnelle d'une information sur l'échelle, pour pouvoir extraire des points d'intérêt d'images à différentes échelles et résolutions.

Les détecteurs de points d'intérêt appartiennent à trois catégories principales :

1. détecteurs de contours : parmi lesquels on peut citer le détecteur de contour de Canny [Canny 86] et Canny-Deriche [Deriche 87], les filtres différentiels, les opérateurs de Sobel, Prewitt, Roberts et Cross ;
2. détecteurs de coins : les détecteurs de Moravec [Moravec 79] et de Harris [Harris 88a] (et leurs différentes améliorations pour les rendre robustes aux changements d'échelles [Lindeberg 98]), aux transformations affines [Shi 94], invariant aux rotations [Harris 88b] qui extraient les coins à partir des dérivées des images. On peut également citer deux approches rapides [Rosten 10] qui déterminent les coins à partir des variations de niveaux de gris de l'image autour d'un noyau ;
3. détecteurs de régions d'intérêt : les régions d'intérêt permettent d'extraire des régions, et leur agencement, contrairement aux détecteurs de coins. Cependant, de nombreux détecteurs de régions d'intérêt associent un point maximal aux régions (généralement le centre de gravité) et peuvent être utilisés comme détecteurs de coins. Les approches les plus répandues reposent sur des différences de gaussiennes à différentes échelles [Marr 80], le Laplacien de Gaussiennes [Lindeberg 94], le déterminant de la matrice Hessienne (utilisée dans l'opérateur SURF [Bay 06]) ou sur l'extraction des régions maximales stables [Matas 04].

D'autres détecteurs spécifiques ont également été mis en place pour extraire des éléments particuliers. C'est le cas des travaux présentés par Rusinol dans [Rusinol 07] dans lesquels l'auteur extrait les différents symboles de plans techniques pour les reconnaître. Il applique alors un traitement semi-local puisqu'il calcule les moments de Hu sur chaque polygone. Il utilise l'ambiguïté introduite par les moments de Hu, associée à un système de vote pour déterminer les images à retenir.

Une autre approche sur les méthodes semi-locales est abordée dans [Kauniskangas 99]. Les moments de Zernike et de Fourier-Mellin sont utilisés sur un voisinage des points d'intérêt détectés avec Harris. Il vient également comparer ces deux techniques aux descripteurs SIFT (Scale-Invariant Features Transform). Ces descripteurs sont largement utilisés dans la littérature car robuste et permet d'obtenir les meilleurs taux de reconnaissance dans de nombreux cas d'utilisation (voir [Lowe

99, Choksuriwong 07]). Les approches semi-locales offrent de bien meilleurs résultats par rapport aux approches globales puisqu'elles permettent de décrire ce qui est important dans l'image de manière plus fine. On ne cherche pas à décrire toute l'image mais seulement ce qui caractérise l'image et la discrimine des autres.

4.3 Apprentissage des caractéristiques

Récemment, les approches basées sur le deep learning ont influencé les recherches dans ce domaine de la recherche par un contenu visuel [WAN 2014].. Nous avons vu que dans les approches précédents, les caractéristiques de l'image sont déterminés manuellement. Avec le deep learning, le réseau de neurones convolutionnel permet d'extraire automatiquement ces vecteurs de caractéristiques [WANG 2015]. Le vecteur de caractéristiques est ensuite utilisé pour la recherche de similarité.

Afin de repérer plus finement des éléments présents dans certaines régions de l'image, on peut utiliser une approche du type Bag of Words ou BOW.. L'idée est de clusteriser les vecteurs de caractéristiques, et d'associer chaque caractéristique locale à un « mot visuel ». Dans ce cas là, c'est le vecteur de mots qui est utilisé pour la recherche de similarité [MOHEDANO 2016].

Le choix de la couche pour l'extraction des caractéristiques et celui de la métrique pour la distance de similarité peut donc être délicat. L'idée des «réseaux siamois» consiste à entraîner le réseau neuronal à reconnaître directement des similarités entre des paires d'images [CHOPRA 2005]. On peut étendre ce concept aux réseaux triplés, prenant en compte également les dissimilarités entre paires d'images [GORDO 2016].

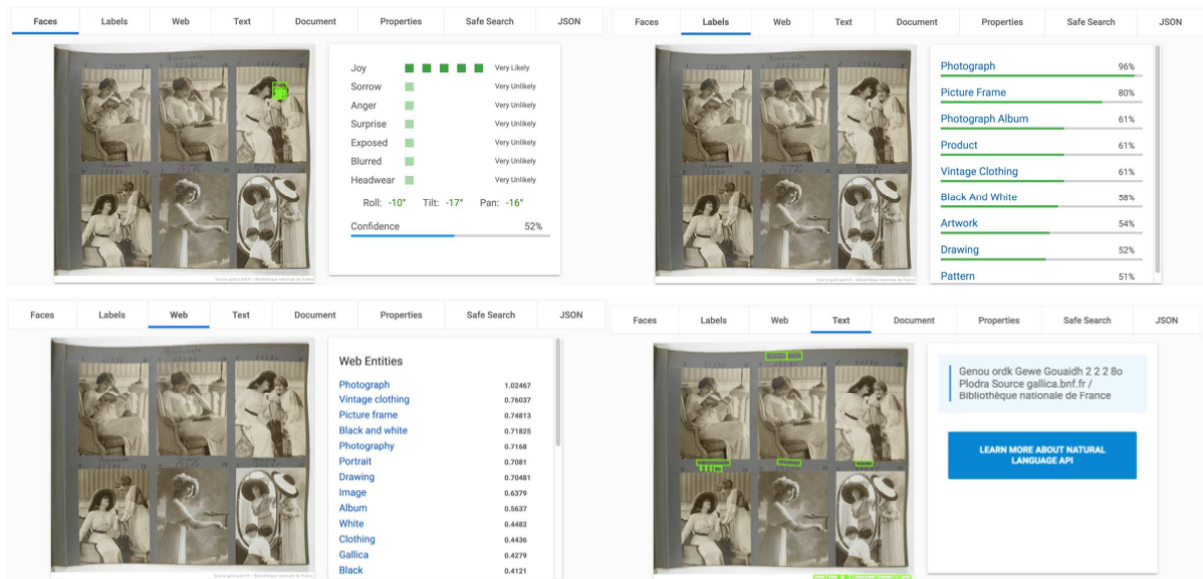
5. Logiciels et boîtes à outils disponibles

Ces dernières années, le déploiement d'applications en vision par ordinateur a été grandement facilitée par la mise à disposition de boîtes à outils, commerciales ou non, permettant un déploiement rapide d'architectures neuronales pour les images. En parallèle, de nombreux modèles ont été rendus publiques et peuvent facilement être utilisés dans des applications locales. Nous donnons ici une liste des outils les plus couramment utilisés. Ce panorama est complété par des pointeurs vers des entreprises françaises capables de développer de telles applications.

Services commerciaux

Les deux grandes applications commerciales permettant de détecter différents concepts dans des images sont respectivement [Google Vision API](#) et [IBM Cloud Visual Recognition](#). Plus récemment, Amazon a également lancé un service similaire, [Amazon Rekognition](#).

Google Vision API. Cette API propose un ensemble de services REST pour la détection d'informations et de concepts dans des images, fournies individuellement ou par lot : extraction et reconnaissance de texte, logos, visages, ou encore différents concepts (appelés *thèmes*) allant "des moyens de transport aux animaux". Un outil de recherche sur le web permet également la recherche d'images similaires et d'attacher des informations de provenance à l'image fournie. En revanche, aucun apprentissage n'est possible.



Exemple d'une image Gallica passée dans Google Vision API. La détection de visage ne semble pas efficace, contrairement à la détection de texte. La détection des web entities se révèlent pertinentes.

IBM Cloud Vision. Comme la précédente, l'API IBM se présente sous la forme de services REST permettant de détecter des concepts dans des images, soumises individuellement ou par lot. Les concepts détectés plus limités que dans l'API Google : visages, ensemble de concepts courants, organisés de manière hiérarchique. En revanche, l'API IBM permet d'apprendre des nouveaux modèles à partir d'exemples négatifs et positifs, fonctionnalité que ne permet pas Google. L'API permet également une recherche d'image par similarité

JSON [↗](#) **PRIVATE BETA**
 Standard Plan members can contact shantenu.agarwal@us.ibm.com to request Private Beta access.

Classes	Score
person	0.60
Museum (Indoor)	0.76
art gallery	0.52
indoors	0.52
silverpoint	0.50
alabaster color	1.00

Type Hierarchy
 /indoors/art gallery

Did We Wow You? Yes No

Text	Score
222	0.94
source	0.98
rationale	0.92
de	0.86
france	0.64

Did We Wow You? Yes No

dans une collection.

exemple de sortie IBM Vision en utilisant la même image que celle donnée à Google.

Amazon Rekognition. Fondée sur le même principe que ses concurrents, Amazon Rekognition exploite la cloud Amazon S3 pour proposer un ensemble de services d'analyse d'images : reconnaissance de concepts / objets, détection et reconnaissance de visages (dont des célébrités), détection de texte dans les images. Les services Amazon s'appliquent aussi à des vidéos. Arrivé après ces concurrents sur le marché, Amazon Rekognition paraît gagner en popularité. Une version démo est disponible après inscription au service et permet une analyse de 5000 images par mois.

Ces trois services principaux sont particulièrement adaptés aux images de scènes naturelles et, comme illustré dans le premier exemple ci-dessus, ne donnent pas toujours des résultats satisfaisants sur des images d'archives (notamment, images noir et blanc, dessins). À noter que seul IBM Cloud Vision permet l'apprentissage pour la détection de nouvelles classes à partir d'exemple annotés, une fonctionnalité qui peut être délicate à mettre en oeuvre à partir des boîtes à outils standards (*cf.* ci-dessous). Enfin, le mode de fonctionnement par service REST permet difficilement l'indexation et la recherche par similarité au sein d'une même collection.

En parallèle des services en ligne, il existe en France plusieurs entreprises sur le créneau de la reconnaissance visuelle et de la recherche d'images. Les deux plus pertinentes pour cette étude sont deux jeune entreprises, [Lamark](#) et [Neovision](#). La première, start-up issue d'Inria Rennes, développe des solutions logiciels autour du tatouage d'images, de la recherche d'images similaires et de la détection d'objets. En combinant ses trois technologies, elle s'impose aujourd'hui comme un acteur de premier plan sur le thème de la protection des droits (copyright notamment). Neovision est quant à elle une société de service en intelligence artificielle qui s'appuie sur un vivier de laboratoires grenoblois. Elle a notamment réalisé des applications en vision par ordinateur, par exemple en reconnaissance de texte ou en recherche d'images similaires. Hors de France, notons également la société [Videntifier](#) qui commercialise une solution de recherche d'images ou de vidéos similaires fonctionnant à très large échelle. Il existe évidemment en Europe et de part le monde plusieurs entreprises similaires à celles mentionnées ici, *e.g.*, la société coréenne [OddConcepts](#), mais l'évolution extrêmement rapide du marché et la domination écrasante des GAFAs rend leur identification périlleuse et peu fiable dans la durée.


Frameworks et réseaux pré-entraînés pour la classification

Il existe aujourd'hui de nombreux *frameworks* pour l'apprentissage profond, pour lesquels il est facile d'obtenir des modèles pré-entraînés que l'on peut appliquer sur les images d'une collection telle celle de la BNF. Ces *frameworks* permettent tous de tirer le meilleur parti des architectures mixtes CPU/GPU, l'utilisation de GPU améliorant considérablement le temps de traitement pour des réseaux profonds de convolution qui font par essence appel à de nombreux calculs tensoriels. Récemment, Google a introduit des nouvelles unités de traitement, les *Tensor Processing Units* (TPU), dédiée à l'apprentissage neuronal profond et qui offre de meilleures performances que les GPU. Si la plupart des *frameworks*, sinon tous, tirent parti de GPU en s'appuyant sur CUDA, seul Tensorflow permet aujourd'hui d'exploiter des TPU. Il y a fort à parier que, dans un avenir proche, d'autres *frameworks* permettront d'exploiter des TPU et que de nouvelles architectures matérielles verront le jour pour accélérer les calculs (*e.g.*, exploitant des FPGA, des calculs flottants approchés, *etc.*). Cependant, il convient de souligner que les opérations coûteuses sont généralement liées à l'apprentissage de nouveaux réseaux, l'utilisation de réseaux pré-existants restant tout à fait abordable en terme de temps de calcul dès lors que l'on dispose de quelques GPU, par exemple les cartes dédiées Titan-X ou Titan-V. Enfin, on trouve typiquement les modèles classiques pour la classification (AlexNet, VGG, Inception, *etc.*) ou pour la détection (RCNN, Fast RCNN, *etc.*) pour les différents *frameworks*, qui s'accompagnent pour la plupart d'un zoo (*sic*) de modèles.

Frameworks classiques

Les *frameworks* les plus populaires, interfacés pour la plupart avec des langages de script, généralement python, et des langages de bas niveau (C, C++) sont les suivants :

Caffe	Framework développé par le Berkeley Vision Lab, sous licence libre (BDS-2). Simple d'utilisation et efficace, caffe propose des interfaces (biding) pour C++, python et matlab. La documentation est cependant plutôt sommaire. Une démonstration permet de tester des modèles classiques, <i>e.g.</i> ,
-----------------------	--

	 <p>CNN took 0.094 seconds.</p>
Tensorflow	<p>Framework open-source développé par Google Brain. Complet et très efficace, il s'accompagne de nombreux tutoriels (officiels ou non). Tensorflow est de loin le framework le plus utilisé bien qu'il soit parfois difficile de trouver des modèles pré-entraînés. Il est utilisé par de nombreux grands comptes, dont naturellement Google. Une interface avec les nouvelles architectures TPU permet un traitement très efficace lorsque ces unités sont disponibles.</p>
pyTorch	<p>pyTorch s'est imposé comme un concurrent sérieux de Tensorflow, notamment utilisé par Facebook et de nombreux académiques. Le <i>framework</i> possède un zoo de modèles pour la classification et la détection facile d'utilisation.</p>
Keras	<p>Keras est une API qui s'appuie sur un <i>framework</i> existant (Tensorflow, theano) de manière à faciliter la programmation de nouveaux modèles en offrant un meilleur niveau d'abstraction et de modularité.</p>

Il existe de nombreux autres frameworks que ceux cités ci-dessus, choisis en raison de leur popularité et de leur chance de survie dans un paysage évoluant à un rythme rapide. Citons par exemple [theano](#), une librairie python ressemblante à keras, ou encore Intel's [neon](#) (zoo fourni !) et [deeplearning4j](#), ce dernier framework étant particulièrement adapté à une distribution Hadoop/Spark pour le déploiement de très grandes architectures. À la date de rédaction de cet étude, il est intéressant de souligner [le tableau comparatif des différents frameworks établis dans wikipedia](#) et que nous reproduisons ci-dessous.

Software	Creator	Software license ^[3]	Open source	Platform	Written in	Interface	OpenMP support	OpenCL support	CUDA support	Automatic differentiation ^[1]	Has pretrained models	Recurrent nets	Convolutional nets	RBM/DBNs	Parallel execution (multi node)
Caffe	Berkeley Vision and Learning Center	BSD license	Yes	Linux, macOS, Windows ^[2]	C++	Python, MATLAB	Yes	Under development ^[3]	Yes	Yes	Yes ^[4]	Yes	Yes	No	?
Caffe2	Facebook	Apache 2.0	Yes	Linux, macOS, Windows ^[5]	C++, Python	Python, MATLAB	Yes	Under development ^[6]	Yes	Yes	Yes ^[7]	Yes	Yes	No	Yes
Deeplearning4j	Skyrim engineering team, Deeplearning4j community, originally Adam Gibson	Apache 2.0	Yes	Linux, macOS, Windows, Android (Cross-platform)	C++, Java	Java, Scala, Clojure, Python (Keras), Kotlin	Yes	On roadmap ^[8]	Yes ^{[9][10]}	Computational Graph	Yes ^[11]	Yes	Yes	Yes	Yes ^[12]
Dlib	Davis King	Boost Software License	Yes	Cross-Platform	C++	C++	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Intel Data Analytics Acceleration Library	Intel	Apache License 2.0	Yes	Linux, macOS, Windows on Intel CPU ^[13]	C++, Python, Java	C++, Python, Java ^[13]	Yes	No	No	Yes	No		Yes		Yes
Intel Math Kernel Library	Intel	Proprietary	No	Linux, macOS, Windows on Intel CPU ^[14]		C ^[15]	Yes ^[16]	No	No	Yes	No	Yes ^[17]	Yes ^[17]		No
Keras	François Chollet	MIT license	Yes	Linux, macOS, Windows	Python	Python, R	Only if using Theano as backend	Under development for the Theano backend (and on roadmap for the TensorFlow backend)	Yes	Yes	Yes ^[18]	Yes	Yes	Yes	Yes ^[19]
MatConvNet	Andrea Vedaldi, Karel Lenc	BSD license	Yes	Windows, Linux ^[20] (macOS via Docker on roadmap)	C++	MATLAB, C++	No	No	Yes	Yes	Yes	Yes	Yes	No	Yes
MATLAB + Neural Network Toolbox	MathWorks	Proprietary	No	Linux, macOS, Windows	C, C++, Java, MATLAB	MATLAB	No	No		No	Yes ^{[22][23]}	Yes ^[22]	Yes ^[22]	No	With Parallel Computing Toolbox ^[24]
Microsoft Cognitive Toolkit	Microsoft Research	MIT license ^[25]	Yes	Windows, Linux ^[26] (macOS via Docker on roadmap)	C++	Python (Keras), C++, Command line ^[26] BrainScript ^[27] (NET on roadmap ^[28])	Yes ^[29]	No	Yes	Yes	Yes ^[30]	Yes ^[31]	Yes ^[31]	No ^[32]	Yes ^[33]
Apache MXNet	Apache Software Foundation	Apache 2.0	Yes	Linux, macOS, Windows, AWS, Android, iOS, JavaScript ^[37]	Small C++ core library	C++, Python, Julia, Matlab, JavaScript, Go, F#, Scala, Perl	Yes	On roadmap ^[38]	Yes	Yes ^[39]	Yes ^[40]	Yes	Yes	Yes	Yes ^[41]
Neural Designer	Artenics	Proprietary	No	Linux, macOS, Windows	C++	Graphical user interface	Yes	No	No	?	?	No	No	No	?
OpenNN	Artenics	GNU LGPL	Yes	Cross-platform	C++	C++	Yes	No	Yes	?	?	No	No	No	?
PaddlePaddle	Baidu PaddlePaddle team	Apache 2.0	Yes	Linux, macOS, Android, Raspberry Pi ^[45]	C++, Go	C/C++, Python	Yes	No	Yes	Yes	Yes ^[44]	Yes	Yes	No	Yes
PyTorch	Adam Paszke, Sam Gross, Soumith Chhriata, Gregory Chanan	BSD license	Yes	Linux, macOS, Windows ^[45]	Python, C, CUDA	Python	Yes		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Apache SINGA	Apache Incubator	Apache 2.0	Yes	Linux, macOS, Windows	C++	Python, C++, Java	No	No	Yes	?	Yes	Yes	Yes	Yes	Yes
TensorFlow	Google Brain team	Apache 2.0	Yes	Linux, macOS, Windows ^[46]	C++, Python	Python (Keras), C/C++, Java, Go, R ^[47]	No	On roadmap ^[48] but already with SYCL ^[49] support	Yes	Yes ^[50]	Yes ^[51]	Yes	Yes	Yes	Yes
Theano	Université de Montréal	BSD license	Yes	Cross-platform	Python	Python (Keras)	Yes	Under development ^[52]	Yes	Yes ^{[53][54]}	Through Lasseigne's model zoo ^[55]	Yes	Yes	Yes	Yes ^[56]

Modèles pré-entraînés : les zoos

La plupart des frameworks possède un zoo de modèles dans lequel on peut trouver les réseaux classiques pré-entraînés. Il est alors possible en quelques lignes de programme de charger ces modèles et de les utiliser pour détecter des classes dans de nouvelles images.

Par exemple, le [zoo associé à caffe](#) propose la plupart des modèles classiques, tels VGG, GoogleNet, ParseNet, ou encore Faster R-CNN. Il est possible de télécharger l'architecture du modèle et les poids (*i.e.*, l'ensemble des paramètres du modèle) pour pouvoir ensuite les utiliser dans une application. On trouve également, en sus des modèles pour la classification, des réseaux pré-entraînés pour générer des signatures compactes d'images

[Lin 2015] qui peuvent être utilisés pour l'indexation et la recherche d'images similaires. Des modèles de segmentation correspondant aux dernières approches sont également disponibles. Il s'agit selon nous du zoo le plus complet parmi tous ceux disponibles.

De manière similaire, on trouve des zoos de modèles pour les autres *frameworks*, e.g., pour [Tensorflow](#) (bien que mal organisé), pour [keras](#) (pas très complet) ou encore pour [pyTorch](#). Ce dernier est bien organisé et facile de prise en main.

De manière intéressante, il est possible pour la plupart de ces *frameworks*, notamment pyTorch, keras ou encore Tensorflow, de récupérer directement des jeux de données standard tels ILVCSR, COCO ou MNIST afin d'évaluer les modèles facilement.

Soulignons, pour finir cet inventaire, la difficulté de convertir un modèle d'un *framework* à un autre qui oblige souvent à installer et maintenir plusieurs frameworks actifs sur une même machine afin de pouvoir facilement bénéficier de la richesse des différents zoos de modèles.

Logiciels d'indexation pour la recherche d'images similaires

Pour la recherche d'images similaires, il n'existe pas vraiment de solutions "sur étagère". Les réseaux pré-entraînés ou encore les bibliothèques standards de vision, notamment [OpenCV](#), permettent d'extraire des descripteurs locaux ou globaux des images. Il reste cependant à indexer ces descripteurs de manière à effectuer des comparaisons de manière très efficace et, dans le cas de descripteurs locaux, à s'assurer d'un minimum de contraintes géométriques dans l'appariement des descripteurs locaux de deux images. Certains réseaux profonds permettent, comme indiqué précédemment, de générer des codes de hachage binaire pour indexer les images. On peut aussi utiliser des logiciels libres d'indexation permettant de rechercher de manière rapide et efficace les proches voisins d'un point dans un espace de grande dimension. Si deux techniques existent pour ce faire (indexation en mémoire primaire ou indexation sur disque), les logiciels libres existants reposent tous sur la première solution et sont donc, à ce titre, sensibles aux crashes (pas de propriétés ACID comme dans une base de données) et gourmands en mémoire vive. Les principaux logiciels d'indexation de vecteurs en grande dimension sont :

FLANN	FLANN est une bibliothèque proposant plusieurs techniques d'indexation pour la recherche approximative de plus proches voisins dans des espaces de grandes dimensions. Elle
-----------------------	---

	propose plusieurs structures d'indexation classiques (e.g., k-means, KD tree, LSH) et des interfaces python, C, C++ et matlab.
Faiss	Faiss est la librairie développée par Facebook pour la recherche approximative de plus proches voisins dans des espaces de grandes dimensions. Fondée sur la notion de <i>product quantization</i> , elle propose des solutions pour les différentes étapes de l'indexation : pré-traitement des vecteurs, partitionnement des données, génération de l'index, recherche des plus proches voisins.
Flickr similarity search	Flickr similarity search est l'équivalent Yahoo! de Faiss et repose sur une amélioration de l'indexation par <i>product quantization</i> susceptible d'offrir des meilleurs propriétés des indexes. Il s'agit d'un ensemble de scripts python permettant de réaliser les différentes étapes de l'indexation et de la recherche.
Yael	Yael est une librairie interfacée en C, python et matlab pour la recherche de plus proches voisins. Elle propose des implémentations efficaces d'algorithmes de clustering, de recherche de plus proches voisins et de gestion de fichiers inversés.

6. Principaux acteurs

En France, il existe un vivier d'équipes de recherche académiques autour de l'analyse, de la classification et de l'indexation des images et des documents.

Les principales équipes de recherche dans les domaines de l'apprentissage pour la classification et l'indexation d'images et l'indexation par le contenu sont (par ordre alphabétique du responsable d'équipe) :

Laboratoire Équipe Responsable	Recherches dans les domaines évoqués dans ce document
Laboratoire Informatique, Image et Interaction Images et contenus Resp. : Antoine Doucet	traitement du signal ; indexation sémantique d'images ; indexation par le contenu
Laboratoire d'Informatique de Paris 6 Machine Learning and Information Access Resp. : Patrick Gallinari	indexation sémantique d'images ; apprentissage
Institut de Recherche en Informatique et Systèmes Aléatoires et Inria Rennes - Bretagne Atlantique Linkmedia - Linking media fragments Resp. : Guillaume Gravier	indexation sémantique d'images ; indexation par le contenu ; indexation multimodale ; sécurité des systèmes d'indexation multimédia
Laboratoire Bordelais de Recherche en Informatique Image et Son Resp. : Vincent Lepetit	segmentation d'images ; indexation sémantique ; indexation par le contenu
EURECOM Multimedia Information Processing Group Resp. : Bernard Merialdo	analyse, traitement, indexation et filtrage de contenus multimédias ; indexation sémantique d'images ; indexation par le contenu
Dpt. Informatique ENS Paris et INRIA Paris WILLOW - modèles de la reconnaissance visuelle d'objets et de scènes Resp. : Jean Ponce__	reconnaissance visuelle d'objets ; compréhension de scènes visuelles ; modèle géométrique et statistiques d'objets, de scènes ; apprentissage
Laboratoire d'Informatique de Grenoble Modélisation et Recherche d'Information Multimédia Resp. : Georges Quénot	indexation et recherche de documents structurés ; indexation sémantique d'images ; recherche d'images similaires

INRIA Rhône-Alpes et Laboratoire Jean Kuntzmann THOTH - Modeling visual knowledge from large-scale data Resp. : Cordelia Schmidt	détection d'objets dans les images ; apprentissage faiblement supervisé ; apprentissage continu ; recherche d'images similaires
Laboratoire GREYC Équipe Image Resp. : David Tschumperlé	traitement d'image et vision par apprentissage profond ; morphologie mathématique ; photographie computationnelle

À cette liste des principaux acteurs académiques pour l'indexation d'images, il convient d'ajouter deux laboratoires privés qui contribuent en France de manière significative à ces domaines :

INA Expert Groupe de recherches audiovisuelles Resp. : Jean Carrive	indexation sémantique ; indexation par le contenu à très grande échelle ; exploration de collections
CEA List Vision and Content Engineering Lab Resp. : Patrick SAYD	indexation sémantique d'images ; indexation par le contenu ; indexation multimodale

Enfin, pour compléter ce premier panorama, mentionnons l'existence de compétences dans des équipes de recherche dont le thème principal n'est pas celui de l'analyse et de l'indexation d'image, comme l'équipe SPARKS du Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis, l'équipe LOGIMAS du Laboratoire de Mathématiques et Informatique pour la Complexité et les Systèmes ou encore l'équipe SaMOVA de l'Institut de Recherche en Informatique de Toulouse.

Dans le domaine de l'analyse de documents, les principales équipes académiques françaises sont :

Laboratoire Équipe Responsable	Recherches dans les domaines évoqués dans ce document
Institut de Recherche en Informatique et Systèmes Aléatoires équipe : Intuidoc Resp. : Éric Anquetil	Communication « homme-document » Analyse automatique de documents numérisés Édition interactive
Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes équipe : DocApp Resp. : Alain Rakotomamonjy	Traitement automatique de l'écrit et des documents Recherche d'Information Bibliothèques numériques
Laboratoire Informatique, Image et Interaction équipe : Images et contenus Resp. : Antoine Doucet	Classification de contenus Indexation/recherche d'informations Traitement d'informations
Laboratoire Informatique Fondamentale et Appliquée de Tours équipe : Reconnaissance de formes et Analyse d'Images Resp. : Nicolas Ragot	Traitement d'images Apprentissage Méthodes interactives
Laboratoire Informatique PARIS DEscartes équipe : Système Intelligent de Perception Resp. : Georges Stammon, Laurent Wendling	Analyse de documents Traitement d'images complexes
Laboratoire d'InforRmatique en Image et Systèmes équipe : Imagine Resp. : Véronique Eglin	Interprétation du contenu Indexation intelligente Reconnaissance et identification d'objets.
Laboratoire Lorain de Recherche en Informatique et ses Applications équipe : QGAR Resp. : Antoine Tabbone	Analyse et reconnaissance de document Reconnaissance de formes.

Notons également la présence de nombreuses équipes de recherche au niveau européen dont les principales sont :

Laboratoire Équipe Responsable	Recherches dans les domaines évoqués dans ce document
CVC, Universitat Autònoma de Barcelona - Espagne équipe : DAG Resp. : Josep Lladós	Analyse de structure, OCR Reconnaissance graphique
PRImA, University of Salford - UK Resp : Apostolos Antanacopoulos	Bibliothèques numériques Évaluation de performance Numérisation, OCR
DFKI - Allemagne équipe : Smart Data & Knowledge Services Resp : Andreas Dengel	Structuration de document Analyse de documents multicanaux
DIVA, University of Fribourg - Suisse Resp : Rolf Ingold	Analyse, Reconnaissance et ingénierie du document
CIL, Institute of Informatics and Telecommunications - Grece équipe : Document processing and understanding Resp : Basilis Gatos	Analyse d'image de documents - OCR Traitement et Analyse de documents patrimoniaux. Analyse de documents Web

7. Conclusion de l'étude.

La BnF désire doter sa bibliothèque numérique Gallica de nouveaux outils d'indexation et de fouilles d'images. Ces outils doivent permettre à tous les utilisateurs (chercheurs, professionnels de la BnF, grand public, ...) une navigation dans les différents corpus grâce à l'utilisation d'une ou plusieurs fonctionnalités permettant :

- une recherche par le contenu : il s'agit là de reconnaître par exemple une lettrine, un visage, une écriture. L'utilisateur présente alors une image et veut retrouver les images identiques ou ressemblantes à son image requête.□
- une recherche par sémantique : ces outils permettent une exploration permettant de rechercher des images correspondant au concept recherché ("véhicule", "visage" ou "poilu" par exemple) ou à une catégorie ("dessin", "photo", "images de telle couleur", ...)
- une recherche hybride, s'appuyant sur les métadonnées catalogue, texte (OCR) et image□

Dans une première partie, le rapport montre les verrous liés à cette problématique :

- Les corpus de la BnF et les documents les composant sont très hétérogènes par leur contenu (différentes époques, nature des supports, techniques d'impression, altérations...) mais également par leur support image (qualité de numérisation, format, compression,...)□
- La volumétrie est très importante et ne comprend pas seulement les documents classifiés comme image, estampe ou photographie mais vise tous les éléments iconographiques qui peuvent être contenus dans un document englobant (par exemple un manuscrit ou un imprimé).□

Afin d'apporter une étude complète sur les outils à mettre en oeuvre pour répondre aux besoins de la BnF, le rapport est alors découpé en plusieurs sections illustrant un état de l'art des technologies, disponibles ou en développement, selon les verrous identifiés :

- En section 3, sur les techniques de prétraitements et de segmentation.□
- En section 4, sur les méthodes d'indexation sémantique.□
- En section 5, sur les méthodes d'indexation par le contenu.□

Pour compléter ce rapport, un panorama de logiciels et boîtes à outils disponibles pouvant répondre aux besoins de la BnF dans leur optique de proposer des outils avancés pour Gallica en section 6. Le rapport présente donc une sélection de services commerciaux, de

frameworks de réseaux profonds (deep learning) classiques ou pré-entraînés et de logiciels d'indexation pour la recherche d'images similaires. Pour finir, il est également proposé une liste des principaux acteurs académiques et équipes de recherche travaillant sur les domaines liés à la problématique en France.

Si le rapport illustre la problématique de la BnF et montre les verrous liés, il avance également un certain nombre de points pouvant y répondre. Il est donc possible d'élaborer certaines hypothèses ou pistes de travail permettant une avancée sur la mise en oeuvre d'outils d'indexation et de fouilles d'images pour Gallica.

Une collaboration menée avec entre la BnF et le L3i a conduit à l'implémentation d'un premier module de recherche dans une collection d'images et d'imprimés relatives à la Première Guerre Mondiale. Cette expérimentation met en avant la pertinence de l'utilisation d'outils basée sur des algorithmes d'apprentissage de type "deep learning". Le prototype développé utilise les fonctionnalités proposées par le service IBM Watson Visual Recognition pour permettre une extraction (et donc une recherche) de concept ou de visages et les conclusions obtenues s'avèrent positives, même si des imperfections sont à corriger. Partant de ce constat, diverses pistes peuvent être envisagées afin de généraliser ces outils à l'ensemble du corpus de Gallica, ou tout du moins une plus grande partie.

La première piste serait bien entendu d'étendre l'utilisation du service IBM Watson VR (ou autre service équivalent) à l'ensemble des collections. Cette solution s'avère pertinente du fait de la maturité et de la disponibilité immédiate des outils. Elle s'accompagne néanmoins d'obstacles qui peuvent rendre la généralisation de leur mise en place délicate dans le cadre de l'application voulue par la BnF. La majorité de ces outils sont en mode "Cloud" ce qui peut induire quelques inconvénients. Leur utilisation nécessite une externalisation des données qui peut être problématique dans le cas de corpus sous droit de diffusion restreint ou privé.

Il est aussi également à rappeler que ces outils nécessitent un apprentissage. Si ces apprentissages permettent ainsi une adaptation au corpus, rien ne peut garantir leur efficacité sur un ensemble aussi conséquent et aussi varié que les banques d'images de la BnF. En effet, ces outils sont notamment développés pour l'exploitation d'images de scènes naturelles et certaines spécificités des images de documents patrimoniaux numérisés ne sont pas pris en considération dans ces développements. Il est donc possible que ces outils conviennent pour une majorité des documents mais atteignent leurs limites sur certains cas (documents dégradés, enluminures spécifiques, gravures anciennes, etc...)

En complément à la première, une seconde piste conduit donc au développement d'outils dédiés en s'appuyant sur les frameworks tels que ceux décrits dans la section 5.2 ou sur des méthodes en développement telles que celles exposées dans les sections 2, 3 et 4. Pour développer ces approches et permettre leur application dans le contexte de cette étude, il conviendrait de maximiser les collaborations entre les laboratoires de recherche (dont certains sont cités en section 6), les services de la BnF et les utilisateurs de Gallica. Ces collaborations ont d'ailleurs toujours fait partie intégrante du développement des services numériques de la BnF. Pour optimiser celles-ci, outre les méthodes éprouvées de collaboration via des réponses à appel à projet de recherche par exemple, deux hypothèses de travail innovantes peuvent être avancées :

- La première consisterait à développer une plateforme permettant l'évaluation des outils que peuvent proposer les laboratoires. La première fonctionnalité de cette plateforme pourrait permettre à la BnF de mettre à disposition un ou plusieurs corpus et d'exprimer ses besoins. Les acteurs participants pourraient alors mener une réflexion sur l'adaptation de leurs méthodes aux données disponibles et au problème puis brancher directement leurs prototypes sur la plateforme (deuxième fonctionnalité), autorisant alors une évaluation conjointe avec la BnF. □
- La deuxième serait fondée sur l'organisation de journées thématiques de type Hackathon. Ces journées viseraient à regrouper un ensemble d'acteurs (BnF, laboratoires informatique et SHS, industriels) pour permettre l'émergence rapide de prototype répondant à des problématiques ciblées au préalable. □

Références Bibliographiques

[Alheritiere 2016] Alhériitière H., Cloppet F., Kurtz C., Vincent N., Utilisation de la couleur pour l'extraction de tableaux dans des images de documents. In Colloque International Francophone sur l'Écrit et le Document 2016 (CIFED), 2016.

[Alheritiere 2017] Alhériitière H., Cloppet F., Kurtz C., Ogier J.M., Vincent N., A document straight line based segmentation for complex layout extraction, In Proceedings of the 14th International Conference on Document Analysis and Recognition - ICDAR2017, p. 1126–1131, 2017.

[Antanacopoulos 1998] A. Antonacopoulos, Page segmentation using the description of the background, Computer Vision and Image Understanding, 70(3), p. 350–369, 1998.

[Antanacopoulos 2013] A. Antonacopoulos, C. Clausner, C Papadopoulos, S. Pletschacher, ICDAR 2013 competition on historical newspaper layout analysis (HNLA 2013). In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), p. 1454–1458, 2013.

[Antanacopoulos 2015] A. Antonacopoulos, C. Clausner, C Papadopoulos, S. Pletschacher, ICDAR 2015 competition on recognition of documents with complex layouts-rdcl2015. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), p. 1151–1155, 2015.

[Antanacopoulos 2017] C. Clausner, A. Antonacopoulos, S. Pletschacher, ICDAR 2017 competition on recognition of documents with complex layouts-rdcl2017. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), p. 1404–1410, 2017.

[Awad 2016] Awad, G., Snoek, C., Smeaton, A. F., Quénot, G., 2016. TRECVID semantic indexing of video : A 6-year retrospective. ITE Transactions on Media Technology and Applications 4 (2016).

[Baird 2007] H. S. Baird, M. A. Moll, Chang An, M. R. Casey, Document image content inventories. In Document Recognition and Retrieval XIV, DRR 2007, San Jose, California, USA, January28, 2007.

[Bay 2006] Herbert Bay, Tinne Tuytelaars & Luc Van Gool. Surf : Speeded up robust features. In European Conference on Computer Vision, pages 404–417, 2006.

[Belaïd 1997] A. Belaïd, Analyse de document : de l'image à la représentation par les normes de codage. Document numérique 1(1), p. 21–38, 1997.

[Beusekom 2010] J.Beusekom, F.Shafait,, T.M. Breuel, Combined orientation and skew detection using geometric text-line modeling. International Journal of Document Analysis and Recognition (IJ DAR) 13(2), p. 79–92, 18, 2010.

[Bloomberg 1991] D. S Bloomberg, Multiresolution Morphological Approach to Document Image Analysis. In International Conference of Document Analysis and Recognition, p. 963–971, 1991.

- [Brown 2007] Brown, M., Sun, M., Yang, R., Yun, L., Seales, W., Restoring 2D Content from Distorted Documents. *Pattern Analysis and Machine Intelligence*, 27, 2007.
- [Bunke 2011] Horst Bunke & Kaspar Riesen. Recent advances in graph- based pattern recognition with applications in document ana- lysis. *Pattern Recognition*, vol. 44, no. 5, pages 1057–1067, 2011.
- [Canny 86] J Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pages 679– 698, 1986.
- [Cattoni 1998] R. Cattoni, T. Coianiz, S. Messelodi, C. M. Modena, Geometric layout analysis techniques for document image understanding: a review. *Rap. tech., TC-IRST Technical Report*, 1998.
- [Cesarini 2001] F. Cesarini, S. Marinai, L Sarti, G. Soda, Trainable table location in document images. In *Proceedings of the 16th international conference on pattern recognition*, p. 236-240, 2002.
- [Chan 2017] K-H Chan, S. K. Im, W. Ke, Fast Binarisation with Chebyshev Inequality, *DocEng '17 Proceedings of the 2017 ACM Symposium on Document Engineering*, p. 113-116, 2017.
- [Chen 2007] Chen N., Blostein D., A survey of document image classification: problem statement, classifier architecture and performance evaluation, *International Journal of Document Analysis and Recognition (IJ DAR)*, vol 10, n°1, p.1-16, 2007.
- [Chen 2017] K. Chen, M. Seuret, J. Hennebert, R. Ingold. Convolutional Neural Networks for Page Segmentation of Historical Document Images, In *Proceedings of the 14th International Conference on Document Analysis and Recognition - ICDAR2017*, p. 965–970, 2017.
- [Choksuriwong 07] Anant Choksuriwong. Reconnaissance d'objets dans une image - Application à la biométrie et à la robotique mobile. PhD thesis, Université d'Orléans, 2007
- [Choksuriwong 2005] A. Choksuriwong, H. Laurent & B. Emile. Etude Comparative de Descripteurs Invariants d'Objets. In *ORASIS'05 - Congrès des jeunes chercheurs en vision par ordinateur*, Mai 2005.
- [Chopra 2005] CHOPRA, Sumit, HADSELL, Raia, et LECUN, Yann. Learning a similarity metric discriminatively, with application Lo face verification. In : *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005. p. 539-546*
- [Chowdhury 2010] Pinaki. Chowdhury, Sukhendu. Das, Suranjana. Samanta & Utthara. Mangai. A Survey of Decision Fusion and Feature Fusion Strategies for Pattern Classification. *IETE Technical Review*, vol. 27, no. 4, pages 293–307, 2010.
- [Coppi 2014] D.Coppi, C. Grana, R. Cucchiara, Illustrations segmentation in digitized documents using local correlation features. In *Pushing the Boundaries of the Digital Libraries Field - 10th Italian Research Conference on Digital Libraries, IRCDL 2014, Padua, Italy, January 30-31, 2014*, p. 76–83, 2014.

[Cote 2014] Cote M., Branzan Albu A., Texture sparseness for pixel classification of business document images. *IJDAR*, 17(3), p. 257–273, 2014.

[Coustaty 2008] Mickael Coustaty, Stephanie Guillas, Muriel Visani, Karell Bertet & Jean-Marc Ogier. On the Joint Use of a Structural Signature and a Galois Lattice Classifier for Symbol Recognition. In *Graphics Recognition. Recent Advances and New Opportunities*, volume 5046 of LNCS, pages 61–70. 2008.

[Coustaty 2011] M. Coustaty, K. Bertet, M. Visani & J.-M. Ogier. A New Adaptive Structural Signature for Symbol Recognition by Using a Galois Lattice as a Classifier. *IEEE Transactions on Systems, Man, and Cybernetics, Part B : Cybernetics*, vol. 41, no. 99, pages 1–13, 2011.

[Deng 2009] Deng, J., Dong, W., Socher, R., jia Li, L., Li, K., Fei-fei, L., 2009. Imagenet : A large-scale hierarchical image database. In : *IEEE Conf. on Computer Vision and Pattern Recognition*.

[Deriche 87] Rachid Deriche. Using Canny's criteria to derive a recursively implemented optimal edge detector. *International Journal of Computer Vision*, vol. 1, no. 2, pages 167–187, 1987.

[Faure 2007] C. Faure, N. Vincent, Document Image Analysis for active reading. In *International Workshop on Semantically Aware Document Processing and Indexing SADPI '07, May2007, Montpellier France*, p.7-14, 2007.

[Gaceb 2013] D. Gaceb, F. Le Bourgeois, J. Duong, Adaptive Smart-Binarization Method for Images of Business Documents, *IEEE. Twelfth International Conference on Document Analysis and Recognition (ICDAR 2013)*, Aug 2013, Washington, USA, United States. p. 118-122, 2013.

[Gatos 2006] B. Gatos, I. Pratikakis, and S. Perantonis, Adaptive degraded document image binarization, *Pattern Recognit.*, vol. 39, no. 3, p. 317–327, 2006.

[Gatos 2009] B. Gatos, K. Ntirogiannis, I. Pratikakis, ICDAR 2009 document image binarization contest (DIBCO 2009). In *Proceedings of the Tenth International Conference on Document Analysis and Recognition - ICDAR2009*, p. 1375–1382, 2009.

[Gatos 2017] B. Gatos, K. Ntirogiannis, I. Pratikakis, ICDAR 2017 document image binarization contest (DIBCO 2017). In *Proceedings of the 14th International Conference on Document Analysis and Recognition - ICDAR2017*, p. 1395–1403, 2017.

[Gomez2017] L. Gomez et al. "Self-supervised learning of visual features through embedding images into text topic spaces." *CVPR*, 2017.

[Goodfellow2016] I. Goodfellow et al. "Deep Learning: Adaptive Computation and Machine Learning series". The MIT Press. 2016.

[Gordo2016] A. Gordo et al. "Deep image retrieval: Learning global representations for image search." *ECCV*, 2016.

- [Goyal2017] Goyal et al. "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering". CVPR 2017.
- [Gros 2007] Gros, P. (Ed.), 2007. L'indexation multimédia : description et recherche automatique. Traité IC2. Hermès.
- [Grzejszczak 2012] 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'Écrit et le Document, 2012.
- [Ha 2016] Dai-Ton, Ha, Nguyen Duc-Dung, Le Duc-Hieu, An adaptive over-split and merge algorithm for page segmentation. Pattern Recognition Letters 80, p. 137-143, 2016.
- [Hadjadj 2014] Z. Hadjadj, A. Meziane, M. Cheriet, Y. Cherfa, An Active Contour Based Method for Image Binarization: Application to Degraded Historical Document Images. ICFHR 2014, p. 655-660, 2014.
- [Hadjadj 2016] Z. Hadjadj, A. Meziane, Y. Cherfa, M. Cheriet, I. Setitra, ISauvola: Improved Sauvola's Algorithm for Document Image Binarization. ICIAR 2016, p. 737-745, 2016.
- [Hamrouni 2014] S. Hamrouni, F. Cloppet, N. Vincent, Séparation imprimé-manuscrit par étude de la linéarité et de la régularité du texte. CIFED 2014 Colloque International Francophone sur l'Écrit et le Document, Nancy, France, p. 155–170, 2014.
- [Harris 88a] C. Harris & M. Stephens. In Proceedings of the 4th Alvey Vision Conference, 1988.
- [Harris 88b] C. Harris & M. Stephens. A Combined Corner and Edge Detection. In Proceedings of The Fourth Alvey Vision Conference, pages 147–151, 1988.
- [He2015] K. He, et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." ICCV, 2015.
- [Hedjam 2016] R. Hedjam, H. Ziaei Nafchi, M. Kalacska, M. Cheriet, Influence of Color-to-Gray Conversion on the Performance of Document Image Binarization: Toward a Novel Optimization Problem. IEEE Trans. Image Processing 24(11), p. 3637-3651, 2015.
- [Howe 2013] N. Howe, Document Binarization with Automatic Parameter Tuning, Volume 16 Issue 3, p. 247-258, 2013.
- [Jain 2012] R. Jain, D. S. Doermann, Logo retrieval in document images. In 10th IAPR International Workshop on Document Analysis Systems, DAS 2012, Gold Coast, Queensland, Australia, March 27-29, 2012, p. 135–139, 2012.
- [Journet 2008] Nicholas Journet, Rémy Mullot, Veronique Eglin & Jean-Yves Ramel. Analyse d'Images de Documents Anciens : une Approche Texture. Traitement du Signal, vol. 24, no. 6, pages 461–479, 09 2008.

- [Kauniskangas 99] Hannu Kauniskangas. Document Image Retrieval with Improvements in Database Quality. PhD thesis, Oulu, 1999.
- [Khun 2009] K. Khurshid, I. Siddiqi, C. Faure, N. Vincent, Comparison of Niblack inspired Binarization methods for ancient documents, DRR, 2009.
- [Kim 2002] I.-K. Kim, D.-W. Jung, R.-H. Park, Document image binarization based on topographic analysis using a water flow model. *Pattern Recognition*, 35, p. 265–277, 2002.
- [Kise 1998] K. Kise, A.Sato, M. Iwata, Segmentation of page images using the area voronoi diagram *Computer Vision and Image Understanding*, 70(3), p. 370–382, 1998.
- [Kisku 2010] Dakshina Ranjan Kisku, Ajita Rattani, Enrico Grosso & Massimo Tistarelli. Face Identification by SIFT-based Complete Graph Topology. *Computing Research Repository*, 2010.
- [Le 2012] V. P. Le, M. Visani, D. C. Tran, J.-M. Ogier, Logo spotting for document categorization. In *Proceedings of the 21st International Conference on Pattern Recognition, ICPR 2012, Tsukuba, Japan, November 11-15, 2012*, pages 3484–3487, 2012.
- [Le Bodic 2009] Pierre Le Bodic, Hervé Locteau, Sébastien Adam, Pierre Héroux, Yves Lecourtier & Arnaud Knippel. Symbol Detection Using Region Adjacency Graphs and Integer Linear Programming. In *10th International Conference on Document Analysis and Recognition*, pages 1320–1324. IEEE Computer Society, 2009.
- [Leydier 2004] Y. Leydier, F. Le Bourgeois, H. Emptoz, Serialized Unsupervised Classifier for Adaptive Color Image Segmentation: Application to Digitized Ancient Manuscripts, *International Conference on Pattern Recognition, ICPR 2004, Aug 2004, Cambridge, United Kingdom. IEEE*, p.494-497, 2004.
- [Lin 2015] K. Lin, H.-F. Yang, J.-H. Hsiao, C.-S. Chen. Deep Learning of Binary Hash Codes for Fast Image Retrieval. *CVPR 2015 DeepVision workshop*.
- [Liang 2008] Liang, J., De Menthon, D., Doermann, D., Geometric Rectification of Camera Captured Document Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(4), p. 591–605, 2008.
- [Lindeberg 94] Tony Lindeberg. *Scale-space theory in computer vision*. Kluwer Academic Publishers, Norwell, MA, USA, 1994.
- [Lindeberg 98] Tony Lindeberg. Feature Detection with Automatic Scale Selection. *Int. J. Comput. Vision*, vol. 30, no. 2, pages 79–116, 1998.
- [Liu 2016] W. Liu et al. "SSD: Single Shot MultiBox Detector." *ECCV 2016*.
- [Lowe 1999] David G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, pages 1150–1157, Corfu, Greece, September 1999.

- [Luo 2011] Luo, S., Fang, X., Zhao, C., Luo, Y., Text Line Based Correction of Distorted Document Images. In: Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on, p. 1494–1499, 2011.
- [Mao 2003] S. Mao, A. Rosenfeld, T. Kanungo, Document structure analysis algorithms: a literature survey. In Proceedings of the 10th Document Recognition and Retrieval Conference, part of the IS&T-SPIE Electronic Imaging Symposium- DRR2003, p. 197–207, 2003.
- [Marr 80] D. Marr & E. Hildreth. Theory of Edge Detection. Proceedings of the Royal Society of London. Series B, Biological Sciences, vol. 207, no. 1167, pages 187–217, February 1980.
- [Mas 2006] Joan Mas, Gemma Sanchez & Josep Lladós. An Incremental Parser to Recognize Diagram Symbols and Gestures Represented by Adjacency Grammars. In Liu Wenyin & Josep Lladós, editors, Selected papers from GREC'05, volume 3926, pages 243–254. LNCS, 2006.
- [Mas 2010] Joan Mas, Josep Lladós, @ Sánchez & Joaquim Armando Pires Jorge. A syntactic approach based on distortion-tolerant Adjacency Grammars and a spatial-directed parser to interpret sketched diagrams. Pattern Recognition, vol. 43, no. 12, pages 4148–4164, 2010.
- [Matas 04] J. Matas, O. Chum, M. Urban & T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing, vol. 22, no. 10, pages 761 – 767. British Machine Vision Computing 2004.
- [McQueen 1967] J. B. McQueen. Some Methods for classification and Analysis of Multivariate Observations. In Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, 1967
- [Meunier 2005] J.L. Meunier, Optimized xy-cut for determining a page reading order. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), p. 347–351, 2005.
- [MODEANO 2016] MOHEDANO, Eva, MCGUINNESS, Kevin, O'CONNOR, Noel E., et al. Bags of local convolutional features for scalable instance search. In : Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. ACM, 2016. p. 327-331
- [Moghaddam 2012] R. F. Moghaddam and M. Cheriet, AdOtsu: An adaptive and parameterless generalization of Otsu's method for document image binarization, Pattern Recognition., vol. 45, no. 6, p. 2419–2431, 2012.
- [Moghaddam 2015] R. F. Moghaddam, M. Cheriet, A multiple-expert binarization framework for multispectral images. ICDAR 2015, p. 321-325, 2015.
- [Moravec 79] Hans P. Moravec. Visual mapping by a robot rover. In IJ- CAI'79 : Proceedings of the 6th international joint conference on Artificial intelligence, pages 598–600, San Francisco, CA, USA, 1979. Morgan Kaufmann Publishers Inc.
- [Nafchi 2014] H. Z. Nafchi, R. F. Moghaddam, M. Cheriet, Phase-Based Binarization of Ancient Document Images: Model and Applications. IEEE Trans. Image Processing 23(7): 2916-2930, 2014.

- [Nagy 1984a] G. Nagy, C.S. Seth, M. Viswanathan, A prototype document image analysis systems for technical journals. *IEEE Computer*, 25(7) p. 10-22, 1984.
- [Nagy 1984b] G. Nagy, C.S. Seth, Hierarchical Representation of Optically Scanned Documents. In 7th ICPR, p. 347–349, 1984.
- [Nagy 2000] Nagy, G., Twenty years of document image analysis in PAMI. *IEEE Tran. Pattern Anal. Mach. Intell.* 22(1), p. 38–62, 2000.
- [Niblack 1985] W. Niblack, An introduction to digital image processing, Prentice-Hall, Englewood Cliffs, NJ, p. 115–116. 1985.
- [Otsu 1979] N. Otsu, A threshold selection method from grey-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1), p.62–66, 1979.
- [Patel2016] Y. Patel et al. "Dynamic Lexicon Generation for Natural Scene Images." *ECCV Workshops*, 2016.
- [Ramel 2006] J.Y. Ramel, S. Busson, M.L. Demonet, AGORA: The interactive document image analysis tool of the BVH project. In *Proceedings – 2nd International Conference on Document Image Analysis for Libraries, DIAL 2006*, p145-155, 2006.
- [Raveaux 2006] Romain Raveaux, Sebastien Adam, Pierre Heroux & Eric Trupin. Learning graph prototypes for shape recognition. *Computer Vision and Image Understanding*, vol. In Press, Corrected Proof, pages –, 2011
- [Redmon2016] J. Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection." *CVPR 2016*.
- [Rivest-Hénault 2011] D. Rivest-Hénault, R. Farrahi Moghaddam, M. Cheriet, A local linear level set method for the binarization of degraded historical document images, *International Journal on Document Analysis and Recognition*, vol. 15, p. 101-124, April 2011.
- [Rosten 10] Edward Rosten, Reid Porter & Tom Drummond. FASTER and better : A machine learning approach to corner detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, pages 105–119, 2010
- [Rusinol 07] Marçal Rusinol & Josep Lladós. A Region-Based Hashing Approach for Symbol Spotting in Technical Documents. In *Seventh IAPR International Workshop on Graphics Recognition*, pages 41–42, septembre 2007
- [Rusinol 2009] M. Rusiñol. Geometric and Structural-based Symbol Spotting. Application to Focused Retrieval in Graphic Document Collections. PhD thesis, Universitat Autònoma de Barcelona, 2009
- [Rusinol 2010] Marçal Rusiñol & Josep Lladós. Efficient logo retrieval through hashing shape context descriptors. In *Ninth IAPR Workshop on Document Analysis Systems*, pages 215–222. ACM, 2010

[Russakovsky 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. Intl. Journal of Computer Vision 115 (3), pp. 211–252.

[Sauvola 2000] J. Sauvola, M. Pietikäinen, Adaptive document image binarization. Pattern Recognition, 33, p. 225–236, 2000.

[Sehad 2015] Abdenour Sehad, Youcef Chibani, Rachid Hedjam, Mohamed Cheriet, LBP-based degraded document image binarization. IPTA 2015, p. 213-217, 2015.

[Shafait 2010] F. Shafait, R. Smith, Table detection in heterogeneous documents. In Proceedings of the 8th IAPR International Workshop on Document Analysis Systems - DAS '10, p. 65–72, 2010.

[Shafait 2011] F. Shafait, T.M. Breuel, The effect of border noise on the performance of projection based page segmentation methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(4), p. 846-851–66, 2011.

[Shechtman 2007] Eli Shechtman & Michal Irani. Space-time behavior-based correlation-Or-how to tell if two underlying motion fields are similar without computing them ? IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 11, pages 2045–56, November 2007.

[Shi 94] Jianbo Shi & Carlo Tomasi. Good Features to Track. In 1994 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94), pages 593 – 600, 1994.

[Smeulders 2000] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta & Ramesh Jain. Content-Based Image Retrieval at the End of the Early Years. IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, pages 1349–1380, 2000.

[Smith 2009] R. Smith, Hybrid page layout analysis via tab-stop detection. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), p. 241–245, 2009.

[Snoek 2005] Snoek, C. G. M., Worring, M., 2005. Multimodal video indexing : A review of the state-of-the-art. Multimedia Tools and Applications 25 (1), pp. 5–35.

[Su 2010] B. Su, S. Lu, and C. L. Tan, A self-training learning document binarization framework, in Proc. 20th ICPR, Aug. 2010, p. 3187–3190, 2010.

[Su 2012] B. Su, S. Lu, and C. L. Tan, A learning framework for degraded document image binarization using Markov random field, in Proc. 21st ICPR, Nov. 2012, p. 3200–3203, 2012.

[Su 2013] B. Su, S. Lu, and C. L. Tan, Robust document image binarization technique for degraded document images, IEEE Trans. Image Process., vol. 22, no. 4, p. 1408–1417, Apr. 2013.

[Tabone 2006] S. Tabbone, L. Wendling & J.-P. Salmon. A new shape descriptor defined on the radon transform. Comput. Vis. Image Underst., vol. 102, pages 42–51, April 2006

[Takezawa 2017] Takezawa Y., Hasegawa M., Tabbone S., Robust Perspective Rectification of Camera-Captured Document Images, 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 06, no. , p. 27-32, 2017.

[Tensmeyer 2017] Tensmeyer C., Martinez T., Document Image Binarization with Fully Convolutional Neural Networks, In Proceedings of the 14th International Conference on Document Analysis and Recognition - ICDAR2017, p. 99–104, 2017.

[Tómasson 2011] Tómasson, G., Sigurórsson, H., Jónsson, B. , Amsaleg, L., 2011. PhotoCube : effective and efficient multi-dimensional browsing of personal photo collections. In : ACM Intl. Conf. on Multimedia Retrieval.

[Torralba 2008] A. Torralba, R. Fergus & W.T. Freeman. 80 Million Tiny Images : A Large Data Set for Nonparametric Object and Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 11, pages 1958–1970, 2008.

[Trupin 2005] Eric Trupin, La reconnaissance d'images de documents : Un panorama / Document images recognition : A survey, Traitement du Signal, 22, p. 159–190, 2005.

[Uttama 2005] Surapong Uttama, Pierre Loonis, Mathieu Delalandre & Jean- Marc Ogier. Segmentation and Retrieval of Ancient Graphic Documents. In GREC, pages 88–98, 2005.

[Vauthier 2012] J. Vauthier, A. Belaïd, Segmentation et classification des zones d'une page de document. In CIFED, 2012.

[Verleysen 2013] C. Verleysen, C. De Vleeschouwer, Learning and Propagation of Dominant Colors for Fast Video Segmentation, ACIVS, vol 8192, p. 657-668, 2013.

[Vieux 2012] R. Vieux, J-P Domenger. Hierarchical clustering model for pixel-based classification of document images. In Proceedings of the 21st International Conference on Pattern Recognition, ICPR2012, Tsukuba, Japan, November 11-15, 2012, p. 290–293, 2012.

[Vinyals 2015] Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2015. Show and tell : A neural image caption generator. In : IEEE Conf. on Computer Vision and Pattern Recognition. pp. 3156–3164.

[Vukotić 2016] Vukotić, V., Raymond, C., Gravier, G., 2016. Bidirectional joint representation learning with symmetrical deep neural networks for multimodal and crossmodal applications. In : ACM Intl. Conf. on Multimedia Retrieval.

[Wahl 1982] F. M. Wahl, K. Y. Wong, and R. G. Casey, Block segmentation and text extraction in mixed text/image documents, Computer Graphics and Image Processing, 19(1), p. 94, 1982

[Wan 2014] WAN, Ji, WANG, Dayong, HOI, Steven Chu Hong, et al. Deep learning for content-based image retrieval: A comprehensive study. In: Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014. p. 157- 166

[Wang 11] Xiang-Yang Wang, Yong-Jian Yu & Hong-Ying Yang. An effective image retrieval scheme using color, texture and shape features. *Computer Standards & Interfaces*, vol. 33, no. 1, pages 59–68, January 2011.

[WANG 2016] WANG, Huafeng, CAI, Yehe, ZHANG, Yanxiang, et al. Deep Learning for Image Retrieval: What Works and What Doesn't. In : *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*. IEEE, 2015. p. 1576-1583.

[Winder 2011] A. Winder, T. L. Andersen, E. H. Barney Smith, Extending page segmentation algorithms for mixed-layout document processing. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, p. 1245–1249, 2011.

[Yang 2008] Mingqiang Yang, Kidiyo Kpalma & Joseph Ronsin. A Survey of Shape Feature Extraction Techniques. In Peng-Yeng Yin, editeur, *Pattern Recognition*, pages 43–90. 2008.

[Yee 2003] Yee, K.-P., Swearingen, K., Li, K., Hearst, M., 2003. Faceted metadata for image search and browsing. In : *SIG CHI Conf. on Human Factors in Computing Systems*. pp. 401–408.

[Yue 2011] Jun Yue, Zhenbo Li, Lu Liu & Zetian Fu. Content-based image retrieval using color and texture fused features. *Mathematical and Computer Modelling*, vol. 54, no. 3-4, pages 1121 – 1127, 2011.

[Zhang 2002] Dengsheng Zhang & Guojun Lu. Shape-based image retrieval using generic Fourier descriptor. *Signal Processing : Image Communication*, vol. 17, no. 10, pages 825 – 848, 2002.

[Zhang 2008] Zhang, L., Zhang, Y., Tan, C.L., An Improved Physically-Based Method for Geometric Restoration of Distorted Document Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(4), p. 728–734, 2008.

[Zhou2014] B. Zhou et al. "Learning deep features for scene recognition using places database." *NIPS* 2014.