MAD: une plateforme mobile pour l'annotation de document vers la classification

Benjamin Duhtil, Vincent Courboulay, Mickael Coustaty and Jean-Marc Ogier*

* *L3I, University of La Rochelle*
Av Michel Crepeau, 17042 La Rochelle, France
Email: bduthil, vcourbou, mcoustat, jean-marc.ogier @univ-lr.fr

RÉSUMÉ. *Aujourd'hui, l'incroyable explosion de l'acquisition mobile d'images ou de documents parait sans limite. Bien qu'il soit facile de les partager sur les réseaux sociaux ou sur le cloud, il est encore très difficile de les classer automatiquement, de les trier ou de rechercher à l'intérieur de cette base de connaissance. Pour répondre à ce défi, nous devons d'abord proposer une annotation automatique pertinente pour pouvoir utiliser par la suite une recherche lexicale robuste. C'est l'objectif de cet article que de proposer une plate-forme mobile pour l'annotation de document automatique, claire, efficace et rapide qui permet d'envisager une recherche en texte clair. Notre approche repose sur une nouvelle méthode d'annotation automatique de documents basée sur la reconnaissance de zones saillantes comme le logo. Les résultats montrent la pertinence de l'approche, nous obtenons une moyenne rappel de 80,6 % et une précision de 100%. La précision (100%) met en évidence la robustesse de l'approche que nous proposons.*

ABSTRACT. *Nowadays, the incredible explosion of mobile acquisition of image or documents is unstoppable. While it is easy to share them, it is still very difficult to automatic classify, sort and search inside this knowledge. To answer this challenge, we first have to provide an appropriate automatic annotation to use lexical search robust methods. It is the objective of this paper to propose a mobile platform for clear, efficient and quick automatic document annotation that enable to consider a plain text search. Our approach relies on a new automatic document annotation method based on logo recognition. We evaluate the method on a corpus composed of 1766 administrative documents The results show the relevance of the approach, we obtain an average recall of 80.6% and a precision of 100%. The precision (100%) highlights the robustness of the approach.*

MOTS-CLÉS : *annotation de documents, saillance, terminologie, application.*

CORIA 2014, pp. 223–238, Nancy, 18-21 mars 2014

KEYWORDS: *document annotation, saliency, terminology, application.*

1. Introduction

Aujourd'hui, l'incroyable explosion de l'acquisition mobile d'image ou de documents parait sans limite. Les tablettes et les smartphones sont deux medium essentiels pour capturer ces informations. Photos de vacances, cartes de visite, publicités et articles de journaux sont généralement dématérialisés par ces dispositifs mobiles. Bien qu'il soit facile de les partager sur les réseaux sociaux ou sur le cloud avec des services comme Flickr, Facebook, Pinterest ou GoogleDrive, il est encore très difficile de les classer automatiquement, de les trier ou de rechercher à l'intérieur de cette base de connaissance. Les documents et les images naturelles ont cette différence fondamentale que les premiers sont produits par l'homme pour l'homme. Cette propriété permet de localiser des informations dans les zones saillantes du document numérisé. Ainsi, il n'est pas surprenant de trouver les *conditions générales de vente*, qui souvent pénalisent le client, écrites en petits caractères gris au bas d'une page. Cependant et de façon générale, l'information nécessaire pour trier ou classer un document est souvent mise en avant par l'intermédiaire d'un logo ou d'un élément saillant et facilement reconnaissable. La saillance d'un élément et sa reconnaissance sont deux éléments essentiels pour le classement rapide d'un document. A l'opposé, la recherche d'un document ou d'un ensemble de documents est presque toujours basée sur du texte brut générique comme *téléphone*, *assurance*, *voiture* et il est donc nécessaire de faire une mise en correspondance entre la requête et le document. Cette mise en relation est loin d'être évidente. Si le logo d'une compagnie d'assurance est reconnu, mais si l'utilisateur spécifie simplement *bagnole*, le système pourrait faire-t-il faire la relation entre l'image et son contenu sémantique ?

Pour répondre à ce défi, nous devons d'abord proposer une annotation automatique appropriée afin d'utiliser une recherche lexicale robuste. C'est l'objectif de cet article que de proposer une plate-forme mobile pour l'annotation automatique de document. Plate-forme claire, efficace et rapide qui permet d'envisager une recherche en texte clair. En règle générale, l'analyse d'image sous la forme de vecteurs de caractéristiques extraites et la formation de mots d'annotation sont utilisés par des techniques d'apprentissage pour tenter d'appliquer automatiquement des annotations aux nouvelles images.

Dans cet article, nous présentons dans la section suivante un bref état de l'art concernant l'annotation. Puis en section 3 nous présentons notre cadre de travail. Finalement en section 4 nous présentons les résultats que nous avons obtenus.

2. L'annotation automatique

Classification et annotation image sont les deux problèmes cruciaux de la vision par ordinateur et l'apprentissage de la machine. L'annotation est propre à chaque document et est souvent garante d'une classification performante. En général, deux types d'annotations peuvent être définis :

– annotation de documents par contenu, qui dépend de la compréhension du texte le composant (chiffre, numéro de téléphone, numéro SIRET ...)

– annotation de documents par contexte qui dépend de la compréhension de la structure (logo, tableau, tampon, stabilo...)

C'est ce dernier type d'annotation auquel nous nous sommes intéressés dans cet article. De nombreux travaux antérieurs ont exploré l'utilisation des caractéristiques globales de l'image pour la classification (Li et Fei-Fei, 2007 ; Oliva et Torralba, 2001 ; Szummer et Picard, 1998 ; Vailaya *et al.*, 2001 ; Vogel et Schiele, 2004). Dans (Li et Fei-Fei, 2007), ils catégorisent des événements sportifs sur des images statiques en intégrant une catégorisation de la scène et des objets. Ils proposent ainsi un modèle d'intégration qui apprend à classer les images statiques sur les événements sociaux complexes. Oliva introduit dans (Oliva et Torralba, 2001) un modèle computationnel de la reconnaissance des scènes naturelles qui n'utilise pas la segmentation et le traitement des objets ou des régions. Les auteurs ont proposé un ensemble de dimensions perceptuelles (naturelle, ouverture, rugosité, expansion...) qui représente la structure spatiale d'une scène. Ces propriétés sont liées à la forme obtenue dans l'espace généré et ont une signification pour les observateurs humains. Cette technique est adaptée pour des images naturelles, en effet le modèle d'enveloppe spatiale organise les photos de scène comme des sujets humains le font, ce modèle de plus est capable de récupérer des images qui partagent la même catégorie sémantique. Dans (Szummer et Picard, 1998) les auteurs ont annoté les scènes en intérieur/extérieur en utilisant des propriétés de haut niveau qui peuvent être déduites à partir de la classification des caractéristiques bas-niveaux de l'image. Vogel a proposé dans (Vogel et Schiele, 2004) une approche d'annotation des scènes naturelles basée sur une mesure de la typicité sémantique. La mesure de la typicité proposée permet d'évaluer la similitude d'une image par rapport à une catégorie de la scène.

Dans les méthodes précédentes, des techniques discriminantes et générative ont été appliquées à ce problème. Les méthodes discriminantes incluent le travail de (Chen et Wang, 2004 ; Lazebnik *et al.*, 2006 ; Zhou et Zhang, 2007). Les méthodes génératives comprennent le travail de (Cao et Fei-Fei, 2007 ; Fei-Fei et Perona, 2005 ; Li et Fei-Fei, 2007 ; Quelhas *et al.*, 2005). Pour l'annotation d'image, plusieurs études ont exploré l'utilisation de modèles probabilistes pour apprendre les relations entre les images et les termes d'annotation (Barnard *et al.*, 2003 ; Duygulu *et al.*, 2002 ; Jeon *et al.*, 2003). Une autre technique très intéressante a été proposée par (Wang *et al.*, 2009). Ils ont présenté une représentation parcimonieuse multi-label pour l'extraction de caractéristiques et de classification dans un contexte d'annotation automatique d'images (voir figure 1). Chaque image est tout d'abord codée dans un supervecteur généré par un modèle de mélange de Gaussiennes sur les patchs d'image. Ensuite, un codage par représentation parcimonieuse permet de réduire la dimensionnalité. Enfin, ce codage est propagée aux images requêtes.

En dépit de ces travaux, aucune de ces approches ne peut être utilisée pour classer ou annoter automatiquement des documents. En effet, toutes ces techniques utilisent des caractéristiques statistiques pour annoter des images naturelles en diffé-

		
Human Annotation	sky jet plane smoke	sky water ships
MSC Annotation	sky jet plane *flight* smoke	sky water ships *island rocks*
		
Human Annotation	grass cat lion mane	trees building garden fountain
MSC Annotation	grass cat *head* lion mane	*sky* trees building *flowers* garden

Figure 1. *Exemple de comparaison d'annotations par représentation parcimonieuse avec la vérité terrain sur Corel5k (lignes supérieures) et Corel30k (lignes du bas).*

rentes catégories, mais elles sont incapables d'annoter différemment deux types de factures. Comme nous l'avons dit précédemment, la raison en est que les documents sont construits par l'homme pour l'homme. Ils ne basent pas leur information contextuelle sur des caractéristiques comme la rugosité ou un modèle de mélange gaussiens. Comme nous pouvons le voir sur la figure 2, des informations contextuelles sont principalement basées sur l'information saillante (intensité, couleur ou orientation).

Dans la section suivante, nous présentons notre approche de l'annotation automatique de document basée à la fois sur l'information visuelle saillante, mais aussi sur une approche de saillance lexicale.

3. Approche proposée

Notre approche procède en trois étapes résumées dans la figure 3. Premièrement, nous analysons le document pour extraire les zones saillantes que nous appellerons vignettes ou régions d'intérêt. Nous supposons que les logos sont une partie saillante de l'image de part la définition même d'un logo, de ce fait, nous pouvons l'extraire en utilisant une approche basée saillance (cf. section 3.1). Dans une deuxième partie, nous apprenons un vocabulaire (terminologie) associée à chaque vignette du document. Pour ce faire, nous utilisons les API proposées par un moteur de recherche web. Ainsi, nous obtenons depuis le web des informations reliées à chaque information saillante de l'image. Enfin, un jeu de mots clés et associé à chaque vignette. Finalement, sur la base des documents retournés, nous extrayons une terminologie associée au document en utilisant des techniques de fouilles de données.

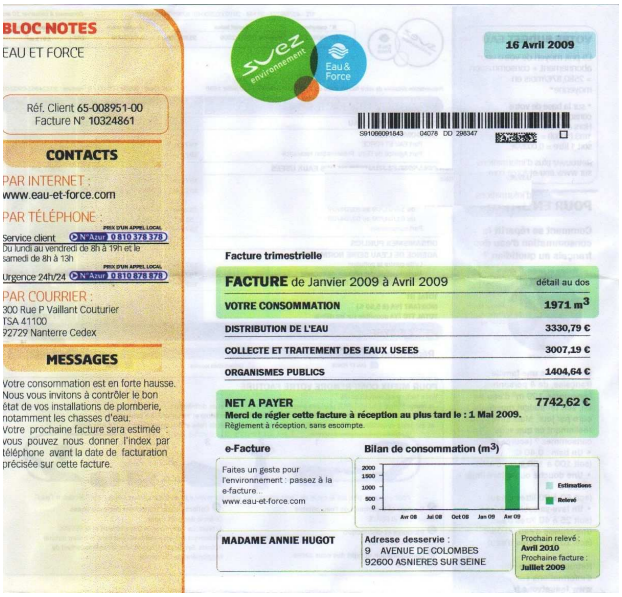

Figure 2. *Exemple de document administratif dans lequel les renseignements importants sont fournis par utilisation de caractéristiques saillantes pour l'œil humain.*

Une de nos motivations étant en effet de réduire le fossé sémantique en proposant une coopération originale entre les techniques bas-niveaux et de fouilles de données haut-niveau.

3.1. *Approche par saillance visuelle*

Le système visuel a inspiré de nombreux modèles d'attention centralisés hiérarchiques, et en particulier les modèles de (Itti *et al.*, 1998) et (Frintrop, 2005). dans ces modèles, la scène visuelle est décomposée en différentes caractéristiques selon une approche multirésolution. Le système génère, pour chacune des caractéristiques prises en compte (intensité, couleur, orientation et mouvement si cela concerne de la vidéo), un certain nombre de cartes représentant les éléments les plus saillants. Les attributs calculés par ce système peuvent être utilisés par le système attentionnel et/ou par un système de vision de plus haut niveau (reconnaissance/suivi d'objets par exemple). Le système attentionnel est principalement compétitif et d'inspiration connexionniste : la compétition entre les différentes cartes de caractéristiques est effectuée par l'interaction de différentes *proies* et *prédateurs* au sein d'un même « écosystème ». On s'éloigne ici des systèmes d'attention hiérarchique pour se rapprocher des systèmes distribués de compétition biaisée.

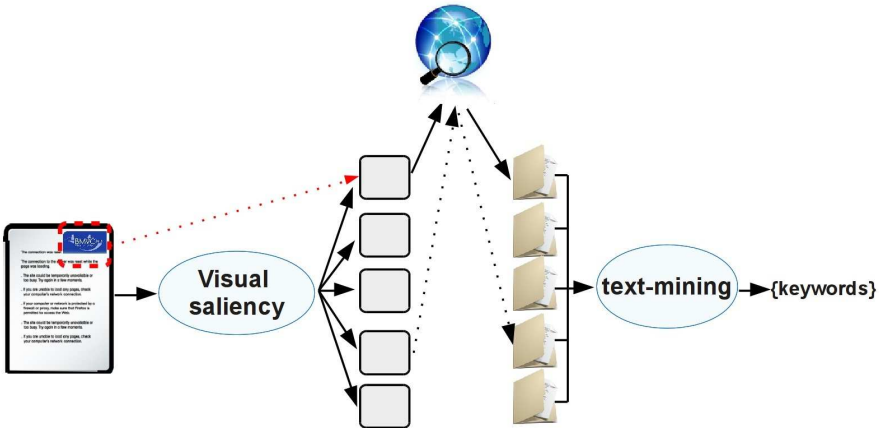

Figure 3. *Shéma d'ensemble de notre approche*

La figure 4 donne un aperçu de l'enchaînement des différents traitements appliqués par notre modèle d'attention. Le vocabulaire utilisé dans ce schéma reprend les termes proposés par (Itti *et al.*, 1998) dans leurs travaux.

De nombreux chercheurs ont travaillé sur la saillance visuelle. Le modèle de Laurent Itti (Itti *et al.*, 1998) est un des plus célèbres. Il est basé sur une approche hiérarchique utilisant une décomposition multi-résolutions et des filtres *centre-périphérie*. Ce modèle fournit une représentation centralisée de l'attention au travers de la génération d'une carte globale de saillance. Le modèle de Le Meur (Le Meur *et al.*, 2006) est une amélioration du modèle d'Itti construit sur la génération de carte de singularité plus réalistes mais aussi plus complexes, ainsi que sur une fusion de cartes de saillance améliorée. Bruce (Bruce et Tsotsos, 2009) propose une approche alternative basée sur la théorie de l'information. Il combine une analyse en composante principale avec une mesure d'information pour estimer la saillance de chaque pixel de l'image. Tous les modèles précédents sont basés sur une représentation centrale de la saillance. Le modèle que nous avons proposé dans (Perreira Da Silva et Courboulay, 2012) fournit une méthode alternative de calcul de l'attention visuelle, basée sur une compétition dynamique des cartes. Cette compétition entre cartes de singularité est pilotée par un système dynamique. Nous utilisons classiquement l'aprroche d'Itti pour extraire les premières cartes de singularité (intensité, couleur et orientation). Ces informations sont proches de celles fournies par la rétine. Ces trois cartes de singularité sont reliées aux mêmes informations que celles que nous activons dans les premières millisecondes lorsque nous regardons un document, une facture ou une publicité. Le second niveau du modèle de Laurent Itti propose un système de fusion des cartes de singularités (C^n) et une simulation du chemin suivi par l'attention vi-

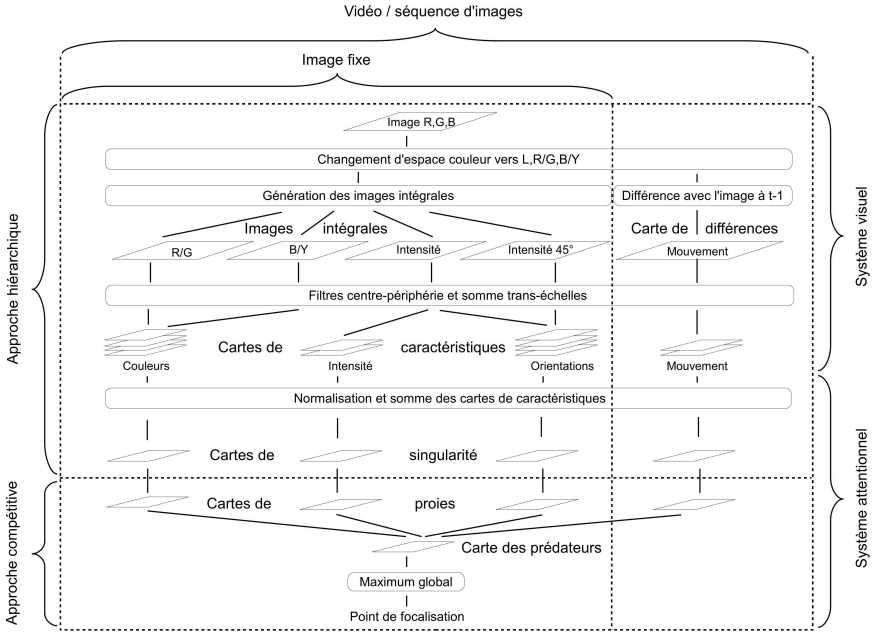

Figure 4. *Schéma général du système d'attention visuelle*

suelle lors d'une observation de scène. Ce focus est déterminé par deux processus de "*winner-takes-all*" et d'algorithmes "inhibition de retour".Nous avons substitué cette seconde partie par notre approche dynamique et compétitive (Perreira Da Silva et Courboulay, 2012), dans laquelle la carte de densité de prédateur I représente le niveau d'intérêt contenu dans une image et les carttes C^n représentent les sources d'intérêt issues des cartes d'intensité, couleur et orientation Afin de simuler l'évolution du focus d'attention, nous proposons un système prédateur-proie (comme décrit ci-dessus) possédant les caractéristiques suivantes :

– le système comprend un type de prédateur et trois types de proies

– ces trois types de proies représentent la répartition spatiale de la curiosité générée par nos trois types de cartes de singularité (intensité, couleur, orientation) ;

– la densité des prédateurs représente l'intérêt. il est généré par la consommation de la curiosité (proies) associé aux différentes cartes de singularité ;

– le maximum global de la carte de prédateurs (intérêts) représente le focus attentionel au temps t.

Pour chacune des cartes de visibilité (couleur, l'orientation de l'intensité), l'évoluation de la population des proies C^n est régie par les équations suivantes :

$$\frac{d_{C^n_{x,y}}}{dt} = C^n_{x,y} + f \triangle_{C^{*n}_{x,y}} - m_C C^n_{x,y} - sC^n_{x,y} I_{x,y} \quad [1]$$

où $C^{*n}_{x,y} = C^n_{x,y} + w{C^n_{x,y}}^2$ et $n \in \{c, i, o\}$, Cela signifie que cette équation est valable pour C^c, C^i andC^o qui représentent respectivement la couleur, l'intensité et l'orientation. w est un paramètres de contrôle de rétroaction. m^n_C est le taux de mortalité qui permet de diminuer le niveau d'intérêt des régions sur la carte de singularité C^n.

La population de prédateurs I, qui consomment les trois types de proies, est régie par l'équation suivante :

$$\frac{d_{I_{x,y}}}{dt} = s(P_{x,y} + wI^2_{x,y}) + sf \triangle_{P_{x,y}+wI^2_{x,y}} - m_I I_{x,y} \quad [2]$$

avec $P_{x,y} = \sum_{n\in\{c,i,o\}} (C^n_{x,y}) I_{x,y}$.

Cela mène aux équations suivant, qui régissent sur une carte bidimensionnelle l'évolution des populations de proies et de prédateurs :

$$\begin{cases} \frac{d_{C^i_{x,y}}}{dt} & = bC^i_{x,y} + f \triangle_{C^i_{x,y}} - m_{C^i} C^i_{x,y} - sC^i_{x,y} I_{x,y} \\ \frac{d_{I_{x,y}}}{dt} & = sC^i_{x,y} I_{x,y} + sf \triangle_{P_{x,y}} - m_I I_{x,y} \end{cases} \quad [3]$$

Le facteur w de retroaction positive assure l'évolution du système et facilite l'émergence de comportements chaotiques en accélérant la saturation de certaines zones de la carte (Perreira Da Silva et Courboulay, 2012 ; Perreira Da Silva *et al.*, 2010 ; Perreira Da Silva *et al.*, 2011).

3.2. *Annotation*

Le processus de fouille de texte est composé des étapes 2 et 3. La la fouille de texte intervient dans l'apprentissage des descripteurs. Cette section présente la méthode de construction automatique du corpus d'apprentissage et d'apprentissage automatique des descripteurs. Nous souhaitons que notre système soit le plus autonome possible (expertise minimale). L'approche *Synopsis* proposée par Duthil et al. (Duthil *et al.*, 2011) entre dans ce cadre. D'une part, *Synopsis* nous permet de construire automatiquement un corpus d'apprentissage à partir de "mots germes", et d'autre part, l'approche nous permet un apprentissage automatique des descripteurs. Dans notre

cas, et uniquement pour la phase de construction automatique du corpus d'apprentissage, "les mots germes" correspondent en réalité aux vignettes saillantes identifiées lors de l'étape 1, c'est pourquoi nous parlerons dans ce cas de "germes visuels". Ces "germes visuels" sont utilisés lors de la phase de construction du corpus d'apprentissage afin de rechercher des documents web contenant le germe visuel considéré (requête par l'image). La notion de "mots germes" n'est cependant pas bannie et elle est utilisée lors de la phase d'apprentissage des descripteurs (mots). Concrètement, les mots germes sont les noms potentiels (noms de logo) des différentes vignettes qui ont été identifiées lors de l'étape 1. Le rôle des mots germes est de guider l'apprentissage des descripteurs textuels : lorsqu'un mot germe est rencontré dans un document, le système apprend les mots qui sont proches, au sens sémantique du terme, du mot germe. Autrement dit, lorsque l'on parle du mot germe, on parle systématiquement d'autres mots sémantiquement proches. Par exemple, lorsque l'on parle de voiture, on pourrait également parler de cylindrée ou de puissance DIN.

3.2.1. *Construction du corpus d'apprentissage*

La construction du corpus d'apprentissage a pour objectif d'obtenir des documents qui ont un contenu similaire à la vignette requête. À chaque vignette saillante est associé un corpus de documents. Dans cette article, nous avons choisi d'utiliser un moteur de recherche web pour construire notre corpus d'apprentissage, d'une part parce que le web est une riche source d'information en perpétuelle évolution (apparition de nouveaux logos, etc), et d'autre part, la diversité des documents disponibles (sites spécialisés, blogs, etc.) nous permet d'apprendre un large panel de descripteurs proches de la vignette requête (registres de langue différents, expressions, etc.). La figure 3 illustre le principe de construction automatique du corpus d'apprentissage.

Plus formellement, à chaque vignette q, q variant de 1 à k, k étant le nombre de vignettes identifiées, un corpus Doc^q de n documents est associé tel que $Doc^q = doc_n^q, n = 1 \dots n^q$.

3.2.2. *Apprentissage des descripteurs*

L'objectif de l'apprentissage est de construire un lexique de descripteurs textuels (mots) décrivant sémantiquement chacune des vignettes identifiées (logos). C'est à dire qu'à chaque logo est associé un lexique L^q. Pour faire le lien entre les éléments graphiques contenus dans le document et les différents descripteurs textuels auxquels ils font référence, nous utilisons le corpus d'apprentissage constitué à l'étape 2. De part sa construction, le corpus d'apprentissage nous permet de faire le lien entre une vignette (image) et la sémantique (page web) qu'elle contient. Ainsi nous sommes en mesure de réduire le fossé sémantique entre le logo et les différents descripteurs textuels auxquels il fait référence.
L'approche *Synospis* est principalement basée sur deux éléments clés : la notion de fenêtre et la notion de classe/anti-classe. La fenêtre permet d'effectuer un apprentissage des descripteurs tout en assurant leur cohérence sémantique avec le mot germe (Duthil, 2012). La notion de classe/anti-classe permet de filtrer le bruit web et donc

de gagner en précision sur le logo à caractériser. L'anti-classe correspond à la fréquence des mots non-inclus dans une fenêtre. L'idée est d'apprendre les mots présents dans les fenêtres centrées sur les mots germes (classe), mais également les mots qui ne sont contenus dans aucune des fenêtres (anti-classe). Une fenêtre est construite en considérant uniquement les noms communs et les noms propres qui sont les deux classes grammaticales reconnues comme porteuses de sens (Kleiber, 1996). Plus formellement une fenêtre de taille sz centrée sur un mot germe g pour un document doc est définie par $F(g, sz, doc) = \{m \in doc / d_{NC}^{doc}(g, m) \leq sz\}$ où $d_{NC}^{doc}(g, m)$ est la distance entre le mot germe g et le mot m.

À partir de ces deux notions, le principe général est de calculer la représentativité d'un mot M dans chacune des deux classes (fréquence d'apparition normalisée $\rho(M)$ dans la classe (mots présents dans les fenêtres) et dans l'anti-classe $\overline{\rho}(M)$ (mots en dehors des fenêtres). Plus formellement la représentativité d'un mot M dans chacune des classes est défini tel que :

$$\rho(M) = \sum_{doc} \sum_{\gamma \in \mathcal{O}(g,doc)} |\mathcal{O}(M, F(\gamma, sz, doc))| \quad [4]$$

$$\overline{\rho}(M) = \sum_{doc} |\mathcal{O}(M, \bigcap_{\gamma \in \mathcal{O}(g,doc)} \overline{F}(\gamma, sz, doc))| \quad [5]$$

À partir de la représentativité d'un descripteur dans chacune des classes, il devient possible de déterminer la proximité sémantique du descripteur M considéré en appliquant une formule de discrimination tel que cela est proposé dans (Duthil *et al.*, 2011). Un score $Sc(M)$ est alors attribué à chaque descripteur. Chaque descripteur constitue une entrée du lexique L^q propre au logo q considéré. Le score d'un descripteur est calculé tel que :

$$Sc(M) = f^2(\rho(M), \overline{\rho}(M)) \quad [6]$$

où f est définie tel que :

$$f^2(x, y) = \frac{(x - y)^3}{(x + y)^2} \quad [7]$$

La figure 5 présente deux exemples de lexiques construits à partir de leur logo respectif.

3.2.3. *Annotation des documents*

L'étape d'annotation consiste à rattacher à un document un ensemble de descripteurs textuels. À partir des lexiques de descripteurs appris lors de l'étape 3, il devient possible d'annoter un document avec l'ensemble des concepts (lexiques) qui lui

Logo	extrait du lexique	Logo	extrait du lexique
	AGPM Groupe AGPM contrat habitation assurance risque contrat assurance vie		**Banque Postale** banque client services financement gestion conseiller

Figure 5. *Exemple de descripteurs*

correspond. De plus nous sommes en mesure d'annoter un document en considérant soit le logo qu'il contient, soit le texte qu'il contient ou alors en considérant ces deux dimensions. Extraire le texte contenu dans un document est un processus complexe de Reconnaissance Optique des Caractères (OCR) très chronophage à grande échelle, mais cette contrainte est désormais une étape systématique dans les systèmes d'information actuels. Il est donc envisageable de considérer ces deux entrées (document image, texte issu de l'OCR) dans notre système d'annotation.

– Annoter sémantiquement un document au format image revient à rechercher, et à identifier, les logos qu'il contient afin de lui rattacher les lexiques associés. Une évolution serait de ne pas systématiquement consulter le web pour identifier un logo, mais plutôt de consulter la base de logo local afin d'identifier le logo : si le logo est présent dans la base, cela signifie qu'il a déjà été appris, sinon une recherche sur le web est nécessaire. L'objectif serait de capitaliser cette information.

– Annoter un document à partir de son contenu textuel (résultat d'OCR) consiste à identifier les lexiques (concepts) qui sont en adéquation avec le document. L'approche *Synopsis* nous permet d'identifier (segmentation), à partir d'un lexique, les zones du document qui traitent du concept (logo) considéré. L'approche utilise une fenêtre glissante (Duthil *et al.*, 2011) centrée sur les noms communs pour identifier les segments de textes pertinents. Cette méthode nous permet également d'obtenir, en plus des parties du document qui traitent d'un concept (importance du concept dans le document), de connaître l'intensité ((Duthil, 2012)) du discours. Ces deux dimensions nous permettent une annotation raffinée et nous apportent plusieurs informations : le concept est-il présent dans le document ? si il l'est, qu'elle place il occupe dans le discours (30% du discours par exemple) ? et quand le concept est présent, quelle intensité a-t-il (niveau d'expertise, de précision) ?

À chaque document est associé un fichier xml qui contient un ensemble d'informations sémantiques : lexiques associés (identifiant du lexique correspondant au nom du logo), intensité du discours et l'importance du discours pour chacun des concepts

(lexiques) qui ont été rattachés. La figure 6 illustre le principe d'annotation d'un document.

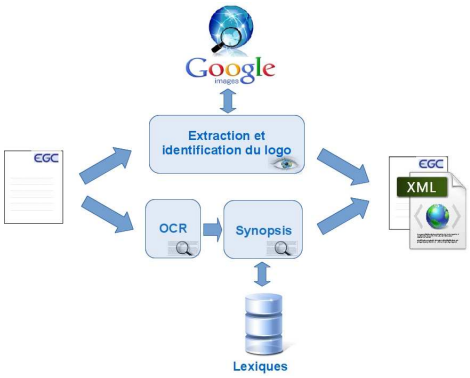

Figure 6. *Processus d'annotation*

4. Expérimentations

Dans cette section nous évaluons la qualité de l'apprentissage sur un corpus de 1766 documents administratifs dans un contexte de classification. Cette base de données a été fourni par un fournisseur de solution de capture de document. Chaque document contient un des 196 logos identifiés dans le corpus. Les documents sont acquis par un **Samsung Galaxy S3** (voir Figure 7).

Ce corpus est composé de 4 classes de documents : acte de mariage (A-M), certificat d'assurance (C-A), relevé d'identité bancaire (RIB) et certificat de naissance (C-N). Le tableau 1 montre la répartition de ces quatre classes (Nombre de logos différents dans la classe et nombre de document de la classe). Cette base confidentielle

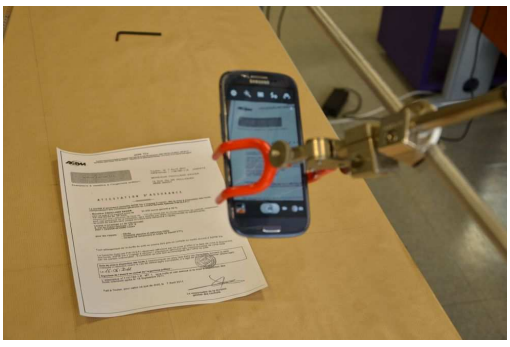

Figure 7. *Système d'acquisition*

provient d'un des leaders mondiaux de la dématérialisation documentaire. Chaque document contient un des 196 logos identifiés dans le corpus. Les documents sont scannés en 200 dpi noir et blanc. Nous utilisons les indicateurs classiques de mesure pour évaluer la classification : *Précision*, *Rappel*. La *Précision* est calculée en considérant les erreurs d'identification du logo : le système identifie un logo qui n'est pas le bon, le lexique associé ne correspond donc pas au logo à identifier. Le *Rappel* est calculé en utilisant le nombre de logos correctement identifié par le système. Nous utilisons durant tout le processus comme moteur de recherche web *Google Images*. Nos résultats sont résumés dans le tableau 2.

	nombre de logos	**nombre de documents**
acte de mariage	36	40
certificat de naissance	113	169
certificat de naissance	30	822
RIB	15	735

Tableau 1. *Répartition des classes*

	C-A	A-M	C-N	RIB	Toutes classes
Rappel	**95,5**	22,5	60,4	**71,7**	***80,6***
Précision	**100**	**100**	**100**	**100**	***100***

Tableau 2. *Résultats de classification pour chaque classe : acte de marriage (A-M), certificat de naissance (C-N), certificat d'assurance (C-A), relevé d'identité bancaire (RIB)*

Les résultats montre la pertinence du système. Les différences de résultats entre chacune des classes s'expliquent par la qualité des documents. Cependant, les résultats sont remarquables, nous obtenons un Rappel de 80,6 et une précision de 100 toutes classes confondues. La Précision met en évidence la robustesse de l'approche. Après cette étude de faisabilité, les tests vont maintenant être complétés avec une étude sur un plus grand nombre de documents puis étendu par l'analyse d'images naturelles (Romberg *et al.*, 2011). Nous avons également entrepris de positionner notre approche vis à vis des méthodes existantes (Romberg *et al.*, 2011).

5. Conclusion and perspectives

Dans cet article nous avons présenté une nouvelle approche d'annotation de documents administratifs. L'approche présentée utilise des méthodes de saillance visuelle et de fouille de texte. L'annotation sémantique qui est proposée permet d'annoter un document sous deux angles : à partir du texte contenu dans le document et/ou du logo qu'il contient. Cette approche permet de réduire le fossé sémantique qu'il existe entre les données bas-niveau contenues dans une image (pixels) et leurs sens. De plus, cette approche s'applique aux documents couleurs et n/b.

Les perspectives sont nombreuses. Tout d'abord, les annotations fournies peuvent être utilisées par un système de recherche d'information. Dans ce cas, il serait nécessaire d'établir une distance sémantique entre les document en considérant l'importance de chacun des concepts présents dans le document (utilisation de l'intensité et de l'importance du concept dans le document) pour être en mesure de retourner les documents les plus pertinents à l'utilisateur en fonction de sa requête.
Dans un contexte de classification, les annotations permettent d'attaquer le problème en considérant l'image (logo) du document ou bien d'un point de vue textuel en utilisant les lexiques associés au document. Considérer l'annotation sémantique permettrait de raffiner la classification et ainsi, plutôt que de considérer uniquement comme classe de document le logo d'une entreprise, de pouvoir identifier de nouvelles classes de documents comme par exemple tous les document d'un secteur d'activité.

6. Bibliographie

Barnard K., Duygulu P., Forsyth D., Freitas N. D., Blei D. M., K J., Hofmann T., Poggio T., Shawe-taylor J., « Matching Words and Pictures », , vol. 3, p. 1107–1135, 2003.

Bruce N. D. B., Tsotsos J. K., « Saliency, attention, and visual search : An information theoretic approach », *Journal of Vision*, vol. 9, n° 3, p. 5, 2009.

Cao L., Fei-Fei L., « Spatially Coherent Latent Topic Model for Concurrent Segmentation and Classification of Objects and Scenes », *IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007*, p. 1-8, 2007.

Chen Y., Wang J. Z., « Image Categorization by Learning and Reasoning with Regions », , vol. 5, p. 913–939, December, 2004.

Duthil B., De l'extraction des connaissances à la recommandation, PhD thesis, Université de Montpellier II, France, 2012.

Duthil B., cois Trousset F., Roche M., Dray G., Plantié M., Montmain J., Poncelet P., « Towards an automatic characterization of criteria, DEXA '11 », *Proceedings of the 22nd International Conference on Database and Expert Systems Applications DEXA 2011*, p. 457, 2011.

Duygulu P., Barnard K., Freitas J. F. G. d., Forsyth D. A., « Object Recognition as Machine Translation : Learning a Lexicon for a Fixed Image Vocabulary », *Proceedings of the 7th European Conference on Computer Vision-Part IV*, ECCV '02, Springer-Verlag, London, UK, UK, p. 97–112, 2002.

Fei-Fei L., Perona P., « A Bayesian hierarchical model for learning natural scene categories », *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, vol. 2, p. 524-531 vol. 2, 2005.

Frintrop S., VOCUS :AVisual Attention System for Object Detection and Goal-Directed Search, Phd, University of Bonn, 2005.

Itti L., Koch C., Niebur E., Others, « A model of saliency-based visual attention for rapid scene analysis », *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, n° 11, p. 1254-1259, 1998.

Jeon J., Lavrenko V., Manmatha R., « Automatic image annotation and retrieval using cross-media relevance models », *Proceedings of the 26th annual international ACM SIGIR confe-*

rence on Research and development in informaion retrieval, SIGIR '03, ACM, New York, NY, USA, p. 119–126, 2003.

Kleiber G., « Noms propres et noms communs : un problème de dénomination », *Meta*p. 567-589, 1996.

Lazebnik S., Schmid C., Ponce J., « Beyond Bags of Features : Spatial Pyramid Matching for Recognizing Natural Scene Categories », *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, p. 2169-2178, 2006.

Le Meur O., Le Callet P., Barba D., Thoreau D., « A coherent computational approach to model bottom-up visual attention », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, n° 5, p. 802-817, 2006.

Li L.-J., Fei-Fei L., « What, where and who ? classifying events by scene and object recognition », *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, p. 1–8, 2007.

Oliva A., Torralba A., « Modeling the Shape of the Scene : A Holistic Representation of the Spatial Envelope », , vol. 42, n° 3, p. 145–175, May, 2001.

Perreira Da Silva M., Courboulay V., « Implementation and evaluation of a computational model of attention for computer vision », *Developing and Applying Biologically-Inspired Vision Systems : Interdisciplinary Concepts*, Hershey, Pennsylvania : IGI Global., p. 273-306, August, 2012.

Perreira Da Silva M., Courboulay V., Estraillier P., « Image complexity measure based on visual attention », *IEEE International Conference on Image Processing - ICIP*, Bruxelles, Belgique, p. 3281 - 3284, September, 2011.

Perreira Da Silva M., Courboulay V., Prigent A., Estraillier P., « Evaluation of preys / predators systems for visual attention simulation », *VISAPP 2010 - International Conference on Computer Vision Theory and Applications*, INSTICC, Angers, p. 275-282, 2010.

Quelhas P., Monay F., Odobez J.-M., Gatica-Perez D., Tuytelaars T., Van Gool L., « Modeling scenes with local descriptors and latent aspects », *Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005*, vol. 1, p. 883-890 Vol. 1, 2005.

Romberg S., Pueyo L. G., Lienhart R., van Zwol R., « Scalable Logo Recognition in Real-world Images », *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, ACM, New York, NY, USA, p. 25 :1-25 :8, 2011.

Szummer M., Picard R., « Indoor-outdoor image classification », *, 1998 IEEE International Workshop on Content-Based Access of Image and Video Database, 1998. Proceedings*, p. 42-51, 1998.

Vailaya A., Figueiredo M. A. T., Jain A., Zhang H.-J., « Image classification for content-based indexing », , vol. 10, n° 1, p. 117-130, 2001.

Vogel J., Schiele B., « A semantic typicality measure for natural scene categorization », *Pattern Recognition Symposium, DAGM*, 2004.

Wang C., Yan S., Zhang L., Zhang H.-J., « Multi-label sparse coding for automatic image annotation », *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009*, p. 1643-1650, 2009.

Zhou Z.-h., Zhang M.-l., « Multi-instance multilabel learning with application to scene classification », *In Advances in Neural Information Processing Systems 19*, 2007.