Améliorer la classification de documents par combinaison de descripteurs visuels et textuels

Olivier Augereau* — **Nicholas Journet**** — **Jean-Philippe Domenger****

* *Gestform, 33 Rue François Arago, 33700 Mérignac*
oaugereau@gestform.com
** *Univ. Bordeaux, LaBRI, UMR 5800, F-33400 Talence, France.*
CNRS, LaBRI, UMR 5800, F-33400 Talence, France.

RÉSUMÉ. *La contribution principale de cet article est de proposer une nouvelle méthode de classification des images de documents combinant les caractéristiques textuelles visuelles extraites respectivement avec les techniques des sacs de mots (BoW) et sacs de mots visuels (BoVW). Alors que les tentatives classiques de combinaison telles que celles basées sur le "Borda-Count" aboutissent à des résultats décevants, nous proposons ici une combinaison par apprentissage. Les expériences de cet article ont été réalisées sur une base de données industrielles de* 1925 *images de document. Ces tests révèlent que la combinaison des information améliore significativement les performances de classification. Notre contribution finale est une discussion concernant les réglages des BoW et BoVW dans un contexte industriel.*

ABSTRACT. *The main contribution of this paper is a new method for classifying document images by combining textual and visual features repectively extracted with the Bag of Words (BoW) and the Bag of Visual Words (BoVW) techniques. While previous attempts have been showing disappointing results by combining visual and textual features with the Borda-count technique, we're proposing here a combination through learning approach. The other contribution of this paper are the experiments conducted on a* 1925 *document image industrial database revealing that this fusion scheme significantly improves the classification performances. Our concluding contribution deals with the choosing and tuning BoW/BoVW techniques in an industrial context.*

MOTS-CLÉS : *BoW, BoVW, combinaison texte image, classification, application industrielle*

KEYWORDS: *BoW, BoVW, text and image combination, classification, industrial application*

CIFED 2014, pp. 91–105, Nancy, 18-21 mars 2014

1. Introduction

Le travail présenté dans cet article a lieu dans un contexte industriel. Une entreprise de dématérialisation telle que Gestform[1] numérise plusieurs millions de documents chaque mois. Ces documents sont de natures très diverses : des documents de ressources humaines (papiers d'identité, feuille de paie, formulaires, etc.) ou encore des notes de frais (tickets de transport, reçus de restaurant, facture d'hôtel, etc). En marge de la simple numérisation physique des documents papiers, les clients souhaitent de plus en plus que les images numériques soient indexées selon le type du contenu de l'image. Cette indexation se matérialise par la labellisation de chaque document. L'objectif de cet article est de proposer une solution permettant de regrouper automatiquement les images de documents en fonction d'exemples, en se basant sur leur texte ou leur visuel.

Afin d'être fidèle au contexte industriel, nous avons constitué une base de données de documents qui soit représentative des 100 000 documents numérisés chaque jour par l'entreprise. Parmi les document de la base, certains sont endommagés (coupées, froissées, déchirées, tâchées, etc.), imprimés sur du papier fin avec de l'encre qui s'efface facilement (comme les tickets restaurant).

La figure 1 présente plusieurs exemples d'images de documents extraits à partir d'un processus de numérisation industriel. Par exemple, les factures et les billets "Flunch", "Quick" et "Buffalo" sont imprimés avec des papiers et des encres de faible qualité. D'autres images, telles que les tickets de métro de la classe "RATP" sont petites, tordues ou pliées.

Dans le cadre d'une comparaison des contenus, la qualité des images n'est pas la seule complexité des documents que nous souhaitons classer. Ainsi, certaines images contiennent très peu de mots comme les tickets de caisse, tickets de trains, etc. Certaines classes d'images ont des contenus très similaires d'un exemple à l'autre (et dissimilaires à ceux des autres classes) rendant potentiellement leur regroupement simple. Au contraire, d'autres classes de documents sont composées d'éléments avec des différences visuelles fortes (mise en page, texte, image) présentant des contenus relativement similaires à des images d'autres classes. Ainsi, là où des images composant les classes "Buffalo", "Flunch" et "Quick" ont la particularité d'avoir un contenu peu varié (mais avec peu de contenu), d'autres images telles que celles composant la classe "IBIS" contiennent des factures d'hôtel ayant la particularité de posséder des contenus variés car ils proviennent de différent hôtels.

Un autre cas de figure complexe que nous avons pu observer est celui des images de documents appartenant à des classes différentes présentant de fortes similitudes visuelles. Par exemple, les trois classes "ElsUk", "ELSES" et "ElsFr" correspondent à des formulaires de cases à cocher qui se ressemblent visuellement, cependant la langue diffère (ils sont respectivement en anglais, espagnol et français).

1. www.gestform.com

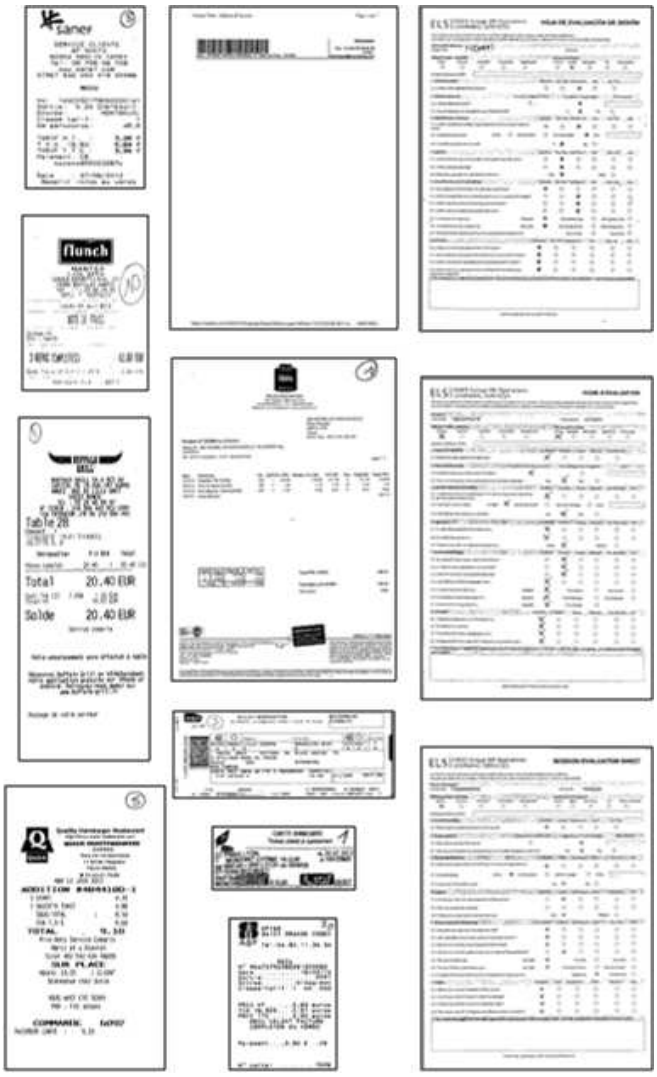

Figure 1. *Exemples de 12 différentes classes de documents. De haut en bas et de gauche à droite : "ElsUk", "ElsEs", "ElsFr", "IBIS", "Lot", "ASF", "SANEF", "SNCF", "Flunch", "Quick", "Buffalo" et "RATP". La qualités des images a été réduite artificiellement pour des raison de confidentialité. Ces documents seront utilisés pour les tests de la dernière partie de cet article.*

Enfin, dans un jeu d'images à analyser, les classes sont généralement composées d'un nombre d'images pouvant être très différent. Certaines classes contiennent peu

de documents telle que la classe "Flunch" tandis que d'autres contiennent beaucoup de document, telle que la classe "Lot".

Figure 2. *Complexité de l'extraction du texte et de la mise en page : deux images originales de la classe "RATP" (ligne 1), les structures extraites (ligne 2) et les résultats de l'OCR (ligne 3). La mise en page et l'OCR sont effectuées par FineReader 10. Pour ces deux documents très similaires, la mise en page et le texte extraits sont très différents.*

Comme nous le verrons dans l'état de l'art, ce genre de problématique a déjà été traité et en grande majorité à l'aide de l'analyse du contenu textuel des images de documents ou l'analyse de la structure (après segmentation). Nos tests en production ont fait très clairement ressortir les limites de l'OCR 2 et des algorithmes de segmentation sur certaines classes de documents. La figure 2 illustre les limites de ces approches. Ainsi, alors que les deux documents (tickets de métro) sont très similaires, la mise en page et le texte extrait sont différents. Dans notre contexte industriel, si certains documents sont facilement "segmentable" ou "océrisable", une grande partie des documents sont dégradés ou fabriqués selon une mise en page complexe à extraire, ce qui rend complexe leur comparaison sur la base de ces informations extraites.

Ce retour d'expérience obtenu en production nous a permis de conclure que l'utilisation de caractéristiques de type "image" pourrait tout à fait être utilisée en complément de celles basées texte dans un contexte de classification de documents.

Dans cet article, nous proposons donc d'appliquer la technique des BoVW pour la classification de documents et de la combiner à la technique des BoW afin d'améliorer les performances de la classification d'images de documents. Nous soulignons l'importance de choisir une technique de fusion pertinente, car les techniques existantes ne permettent pas systématiquement d'obtenir un gain après fusion des deux catégories de caractéristiques. Une autre contribution de cet article est une discussion détaillée sur la façon d'appliquer ce système de classification dans un contexte industriel.

Après la présentation des travaux connexes au sujet, les techniques de BoW et BoVW seront décrites ainsi que les méthodes de combinaison. Enfin, une discussion autours des résultats et de l'utilisation de ces méthodes dans un contexte industriel sera faite.

2. État de l'art

Les auteurs de (Chen et Blostein, 2007) présentent les principales techniques de classification d'images de documents. Trois types de caractéristiques sont distinguées pour la tâche de classification : le texte (mots, chiffres, etc) qui est extrait par un OCR, l'image (couleurs, textures, formes, etc) et la mise en page (description de la structure des documents). Il peut également être souligné que la plupart des techniques présentées dans cet articles sont basées sur un apprentissage supervisé. Nous pensons également que l'apprentissage supervisé peut aider à gérer la complexité car il permet de prendre en compte la diversité des documents au sein d'une classe et la dissemblance des documents entre les classes. Les techniques BoW et BoVW sont systématiquement utilisées avec un apprentissage supervisé.

2.1. *Les sacs de mots*

La technique des BoW est très répandue dès lors que l'on souhaite classer les images de documents selon leur contenu textuel. Elle consiste à résumer un texte à un vecteur qui mesure la fréquence d'apparition d'un ensemble de mots. En 2002, deux principaux travaux ont été publiés au sujet de la classification de texte : l'étude de Sebastiani (2002) et le livre de Joachims (2002). Tout deux mettent en avant que la technique des BoW est l'une des meilleurs méthodes existantes pour la classification de texte automatique.

Les BoW sont utilisés pour de nombreuses tâches telles que les systèmes de recommandation, la classification d'e-mail, l'analyse de sentiments, l'analyse d'opinions (Liu, 2012), etc.

2.2. *Les sacs de mots visuels*

La méthode des BoVW est très utilisée dans la communauté de la vision par ordinateur pour la classification d'images naturelles. Par exemple, (Yang *et al.*, 2007) l'utilise pour la reconnaissance d'objets (Lowe, 2004) et (Valle et Cord, 2009) pour le CBIR (*Content Based Image Retrieval*). Les applications réussies sur des images naturelles ont conduit d'autres communautés à l'utiliser. Cependant, peu d'applications aux images de documents ont été proposées. La technique BoVW a été appliquée à la reconnaissance de logo (Rusinol et Lladós, 2009), de *Word Spotting* (Shekhar et Jawahar, 2012) et à la reconnaissance de caractères manuscrits (Song *et al.*, 2011).

Nous pouvons remarquer que, dans le domaine de l'image du document, la technique BoVW a surtout été appliquée à la recherche de sous-parties d'images : logo, mot, etc. Très peu d'applications utilisant la technique BoVW pour classer l'ensemble d'une image de document ont été proposées. Récemment, les auteurs de (Rusinol *et al.*, 2012) et (Gordo *et al.*, 2013) ont détaillé un système de reconnaissance de documents multi-pages en utilisant des caractéristiques visuelles et/ou textuelles. Ils concluent que les caractéristiques visuelles réduisent les performances de reconnaissance quand ils sont combinés avec des caractéristiques textuelles.

2.3. *Techniques de fusion multimodale*

De nombreuses études ont déjà abordé le sujet du traitement des données multimodales. Par exemple, dans le domaine de l'analyse vidéo, trois modes distincts sont généralement utilisés : le son, l'image et le texte (Snoek et Worring, 2005). Afin de combiner automatiquement les différents modes, deux approches principales se distinguent : la fusion précoce et la fusion tardive. La fusion précoce consiste à trouver un moyen de combiner les différents modes avant la classification. Par exemple, si les caractéristiques sont représentées par des vecteurs, un moyen simple de faire la fusion précoce est de les concaténer. Pour la fusion tardive, les caractéristiques de chaque mode sont utilisées séparément et ensuite leur sorties sont combinées. Une façon simple de faire de la fusion tardive est de faire voter chaque mode.

Selon Meüller (Meüller *et al.*, 2010), la fusion tardive fonctionne généralement mieux que la fusion précoce dans la plupart des tests appliqués à la recherche d'information multimodale basée sur des combinaisons de caractéristiques visuelles et textuelles. Les trois principaux types de fusion tardives sont :

1) La combinaison de scores ou de rangs (Ho *et al.*, 1994) (tel que le *Borda-count*). Les classifieurs renvoient un rang pour chaque classe, puis la combinaison est faite en appliquant un ré-arrangement (*re-ranking*).

2) La combinaison par des règles mathématiques (moyenne, somme, produit, etc.) (Kittler *et al.*, 1998), (Ye *et al.*, 2012). Les classifieurs renvoient une probabilité pour chaque classe, la combinaison est faite en additionnant ou multipliant les probabilités.

3) La combinaison par apprentissage (Gorgevik et Cakmakov, 2005), (Terrades *et al.*, 2009). Chaque classifieur renvoie une mesure pour chaque classe (probabilité, distance, etc). Puis, un autre apprentissage est fait sur ces mesures.

Les auteurs de (Gorgevik et Cakmakov, 2005) ont testé 38 combinaisons de SVM pour la reconnaissance de chiffres manuscrits. Il apparait clairement que la combinaison par apprentissage surpasse la combinaison des rangs ou l'utilisation de règles.

3. Notre mise en œuvre de la classification des documents

3.1. *Classification par sacs de mots*

La technique des BoW repose sur le principe selon lequel un document peut être décrit en comptant les occurrences d'un ensemble prédéfini de mots.

Dans une première étape, on applique un OCR sur les images de documents de manière à extraire les mots. Comme la sortie de l'OCR n'est pas parfaite, un prétraitement est appliqué sur le texte obtenu. Les caractères spéciaux tels que la ponctuation sont supprimés. Nous supprimons également les mots qui contiennent moins de 4 lettres et ceux de plus de 15 lettres. Après cela, trois autres étapes sont généralement appliquées : enlever les *stop words* et appliquer une racinisation et/ou une lemmatisation . Les *stop words* sont des mots tels que des articles ou des pronoms qui apportent peu de sens au texte. La racinisation et la lemmatisation consistent à remplacer les différentes formes d'un mot en un seul terme. Dans le cadre de nos tests, nous n'avons pas utilisé cette étape dans la mesure où nous traitons des documents avec un vocabulaire très peu étendu et contenant peu de phrases (factures, reçus, tickets de caisse, etc). Après le prétraitement, un dictionnaire est créé afin de définir l'ensemble des mots qui seront utilisés pour comparer les documents. Les 1000 mots les plus fréquents sont choisis pour la construction de ce dictionnaire. La caractéristique utilisée pour décrire un document est le nombre d'occurrences de chaque mot du dictionnaire. Finalement, chaque document est décrit par un histogramme de 1000 valeurs . Un SVM est entrainé avec les documents d'apprentissage et les documents de la base de test sont finalement classés.

3.2. *Classification par sacs de mots visuels*

La technique des BoVW repose sur la détection de points d'intérêts tels que SURF (Bay *et al.*, 2008) ou SIFT (Lowe, 2004) qui sont robustes aux transformations telles que la rotation, la translation, la mise à l'échelle, le flou et le changement de luminosité.

Les BoVW sont directement inspirés par les BoW. Dans la plupart des applications, les mots visuels sont des points d'intérêt. L'algorithme BoVW est décrit dans la figure 3. Il peut être résumé en quatre étapes principales. Tout d'abord, les points d'intérêt sont extraits de chaque image à classer. Ensuite, tous ces points sont regroupés dans K groupes. Dans une troisième étape, l'image est décrite par un histogramme de k valeurs. Chaque point de l'image d'intérêt incrémente la valeur de l'histogramme correspondant à son groupe. Enfin, les images sont décrites par leur histogramme qui est utilisé par un algorithme d'apprentissage automatique afin de classer les documents.

Habituellement, les points d'intérêt sont extraits avec SIFT ou SURF, le regroupement des points d'intérêt se fait avec l'algorithme k-means et l'apprentissage supervisé se fait avec SVM (Csurka *et al.*, 2004), (Valle et Cord, 2009). Nous avons utilisé l'im-

plémentation de SURF et de k-means de la bibliothèque OpenCV. Nous avons choisi $k = 1000$ pour le nombre de classes de k-means. Nous utilisons un SVM multiclasse avec un noyau RBF . Les paramètres sont auto-paramétrés par validation croisée en utilisant la bibliothèque libSVM.

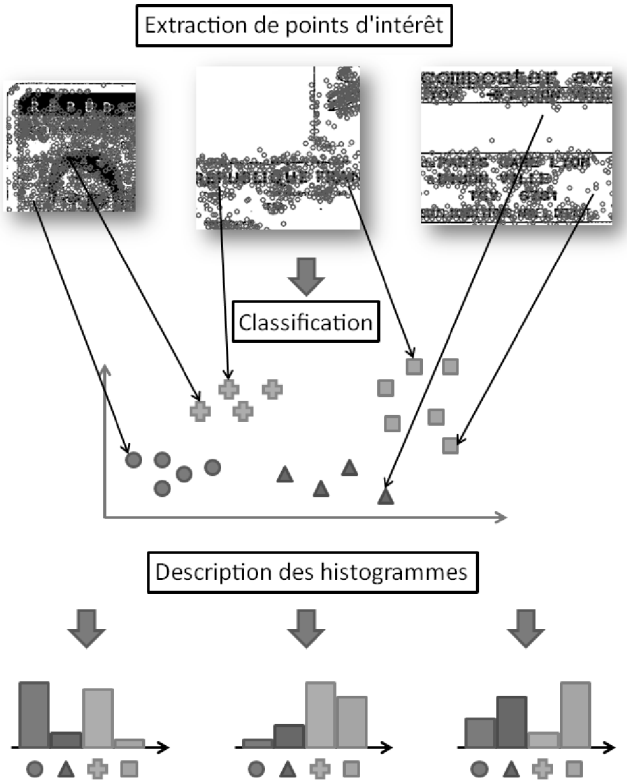

Figure 3. *Description de la création des histogrammes créé par les BoVW. 1) Des points d'intérêt sont extraits sur chaque image. 2) Les points sont regroupés dans k groupes. 3) Chaque point d'intérêt étant lié à l'un des k groupes, chaque image est décrite par un histogramme correspondant au nombre d'occurrences de chaque classe.*

3.3. *Classification par combinaison des BoW et BoVW*

La figure 4 résume la technique de fusion tardive que nous proposons. Les BoW et les BoVW sont appliqués séparément. Par défaut, les SVM ne retournent pas de probabilités. C'est pourquoi nous utilisons de la méthode de Platt (Platt, 1999) qui est l'une des méthode les plus connues pour convertir les sorties de SVM en probabilités. Ensuite, les probabilités sont concaténées, normalisées et fournies en entrée d'une nouvelle étape d'apprentissage (un autre classifieur SVM est utilisé). Les mêmes images

d'apprentissage sont utilisées pour les BoW, les BoVW et l'apprentissage pour la fusion.

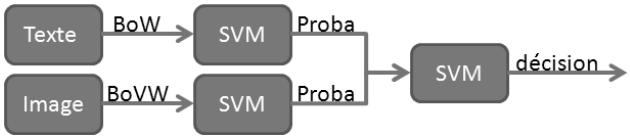

Figure 4. *Schéma de fusion tardive. Les SVM sont appliqués séparément sur les BoW et les BoVW. Les sorties des SVM (probabilités) sont concaténées et fournies en entrée d'un autre SVM.*

4. Tests et résultats

La base de données est composée de 1985 documents numérisés par la société Gestform. Ces documents ont été choisis aléatoirement dans une chaîne de production. Les images sont numérisées à 300 dpi et sont automatiquement binarisées par le scanner. La base de données contient 12 classes de documents. Un exemple de chaque classe est représenté dans la figure 1. Certaines classes de la base de données sont complexes, les détails sont fournis dans l'introduction.

Un travail similaire récent proposé par les auteurs de (Rusinol *et al.*, 2012) a conclu que la fusion des BoW et des BoVW n'était pas pertinente dans le cadre de la recherche d'images de documents. En règle générale, l'ajout des caractéristiques basées sur l'image réduit les performances obtenues uniquement avec le texte. Dans le contexte de classification, la même conclusion est obtenue lors de nos expérimentations.Nous appliquons une fusion tardive en utilisant la combinaison de rangs (technique du *Borda-Count*), les caractéristiques visuelles diminuent les résultats des caractéristiques textuelles. Le tableau 1 montre que la performance baisse de 8 % quand nous combinons les caractéristiques textuelles et visuelles au lieu d'utiliser seulement les caractéristiques textuelles. Nous obtenons donc des résultats cohérents avec ceux trouvés par (Rusinol *et al.*, 2012).

Tableau 1. *Rappel et précision moyennes des techniques de BoW, BoVW et fusion par Borda Count.*

	Rappel	Précision
BoVW	0,871	0,869
BoW	0,982	0,977
fusion Borda	0,899	0,902

Dans la section suivante, les tests de classifications d'images de documents utilisant BoW, BoVW et la fusion par apprentissage sont décrits. Dans le tableau 2, les

résultats de la classification sont détaillés classe par classe pour mettre en évidence les cas où les caractéristiques visuelles conduisent à une meilleure classification que les caractéristiques textuelles. Enfin, les tests montrent que la fusion par apprentissage, contrairement à la fusion *Borda count*, peut produire une signature pertinente pour chaque document en utilisant des informations à la fois visuelles et textuelles.

4.1. *Protocole de test*

Les seuils mentionnés dans cette section ont été choisis lors de nos expérimentations faites en condition réelles. Ils dépendent évidemment de la base testée et sont donc donnés afin d'indiquer au lecteur un ordre de grandeur dans lequel il faut fixer ces paramètres pour une base de 2000 images, 12 classes à identifier dont certaines sont plus simples que d'autres à analyser (cf. introduction). Le choix de la taille de 2000 correspond, en moyenne, à ce qu'un opérateur humain peut labelliser manuellement en quatre heures (une demie-journée de travail). Ce choix des 2000 images permet donc de mettre en perspective l'intérêt du système de classification présenté ici, qui certes produira des erreurs, mais permettra avant tout de faire gagner du temps à un opérateur qui jusque là effectuait la labellisation entièrement manuellement.

Les performances relatives à nos tests sont calculées en utilisant un modèle de validation croisée. Ainsi, pour chaque classe, 5 documents sont choisis au hasard pour l'apprentissage. Tous les 1925 autres documents sont utilisés pour le test. Cette méthode intègre une phase d'apprentissage avec des images choisies aléatoirement. Elle est donc appliquée 10 fois de suite pour palier au biais possible du tirage aléatoire des éléments d'apprentissage. Les résultats exposés (rappel et précision) sont donc issus d'une moyenne des 10 essais.

Habituellement, les méthodes supervisées utilisent environ 60 % de chaque classe pour l'apprentissage et 40 % pour les tests. On peut noter que, dans notre test, seules 5 ont été suffisantes pour l'apprentissage de chaque classe. Ce choix a été fait pour deux raisons liées au contexte industriel. La première est que nous ne connaissons pas la base de données à l'avance, donc nous ne savons pas combien de documents composent chaque classe, il est donc impossible de prendre un pourcentage de chaque classe. La seconde raison est que la classification manuelle (création de la vérité terrain) prend beaucoup de temps. Pour pouvoir être appliquée de manière concrète, il est donc important que cette étape soit réduite autant que possible.

4.2. *Résultats*

Les résultats sont affichés dans le tableau 2. Le nombre de documents par classe utilisé pour les tests est affiché dans la deuxième colonne (les 5 documents par classe utilisés pour l'apprentissage ne sont pas pris en compte).

Les résultats montrent que les BoW ont de très bonnes performances. Cependant, certains documents sont manquants pour les classes tels que "IBIS" (facture d'hôtel)

et "RATP" (tickets de métro). Pour la classe "IBIS", l'explication dans le fait que les documents proviennent de différents hôtels et contiennent des textes parfois différents. De plus, le texte dans le logo et les conditions générales écrites en petits caractères sont difficilement lisibles par un logiciel d'OCR. Pour la classe "RATP", la qualité de l'image est la principale raison impliquant une faible valeur du rappel (figure **??**).

Globalement, les résultats des BoVW sont bons, mais légèrement moins que ceux obtenues à l'aide des BoW. Les trois classes " ELSES ", " ElsFr " et " ElsUk " sont très similaires visuellement et dégradent fortement les performances. On remarque également que pour certaines classes où les BoW ne fonctionnent pas bien (comme "IBIS" ou "RATP") , les BoVW fonctionnent mieux.

La combinaison conduit à deux résultats principaux : 1) l'amélioration soit des BoW, soit des BoVW ou 2) l'amélioration à la fois des BoW et des BoVW . Si une technique (BoW ou BoVW) fonctionne bien mieux que l'autre, la combinaison peut fournir des résultats intermédiaires (cas 1). Par exemple, pour la classe "IBIS", la précision de la fusion est meilleure que la précision des BoW, mais plus mauvaise que la précision des BoVW. L'explication est que pour quelques documents, un classifieur donne la bonne réponse avec un taux de confiance faible alors que l'autre classifieur donne une mauvaise réponse avec un taux de confiance élevé.

Dans la plupart des cas, la technique de fusion fournit des résultats meilleurs ou égaux que l'utilisation individuelle des BoW ou des BoVW. La catégorie "SNCF" illustre un cas intéressant où la fusion surpasse la précision des BoW et des BoVW. Ceci peut être expliqué par le fait que, pour une classe donnée, les deux classifieurs fournissent de bonnes réponses sur les deux sous-parties distinctes de la classe. Enfin, il est également important de garder à l'esprit que le gain des derniers pourcentages en rappel et précision sont les plus difficile à obtenir (amélioration de 98 % à 99 % ou de 99,90 % à 99,99 %).

4.3. *Retour sur l'implémentation industrielle*

L'application de techniques de recherche dans un contexte industriel est complexe. De nombreuses techniques sont *ad hoc* ou basées sur des règles, ce qui les rends difficilement applicables à des contextes hétérogènes. L'utilisation des statistiques textuelles et visuelles permet d'appliquer la classification sur de nombreux types de documents.

Au sujet des paramètres

L'une des principales force des BoW et des BoVW, c'est qu'ils n'ont pas besoin de beaucoup de paramètres pour fonctionner correctement. Il suffit de choisir trois paramètres : le nombre de groupes dans k-means pour les BoVW, la taille du dictionnaire des BoW et le seuil de la matrice Hessienne pour l'extraction des points d'intérêt. Comme énnoncé précédemment, ces seuils ont été fixés pour une base de 2000 images contenant un grand nombre de caractéristiques représentatives de ce qu'un opérateur peut observer sur une demie-journée de travail.

Tableau 2. *Rappel et précision pour les BoW, BoVW et la fusion par apprentissage. Les moyennes sont pondérées par le nombre de documents.*

		Rappel			Précision		
Classes	Nb docs	BoW	BoVW	Fusion	BoW	BoVW	Fusion
SNCF	194	1	0,984	0,995	0,964	0,954	0,990
ElsEs	299	1	0,677	1	0,987	0,709	0,997
ElsFr	99	1	0,816	1	1	0,626	1
ElsUk	430	1	0,782	1	1	0,802	1
IBIS	41	0,810	1	0,972	0,829	0,902	0,854
Lot	670	1	1	1	0,975	0,987	0,991
ASF	31	1	0,792	1	0,742	0,613	0,645
SANEF	22	1	0,167	1	0,955	0,682	1
Buffalo	13	1	1	1	1	1	1
Flunch	4	0,667	0,231	0,400	1	0,750	1
Quick	12	1	0,923	1	1	1	1
RATP	110	0,764	0,940	0,940	1	1	1
Moyenne		0,982	0,872	**0,994**	0,977	0,870	**0,986**

Les deux premiers paramètres ont été fixés à 1000 et le seuil de la matrice Hessienne à 500. Modifier le Hessien a peu d'impact sur les résultats parce que les documents sont binaires. Le nombre de groupes et la taille du dictionnaire ont aussi peu d'impact, l'objectif est d'avoir suffisamment de mots (visuels) pour décrire les documents sans en utiliser trop au risque d'être trop discriminant. Utiliser une valeur comprise entre 200 et 2000 permet d'avoir un fonctionnement correct dans la plupart des cas.

Comment choisir entre les BoW, les BoVW et la fusion

Les tests effectués dans ce document montrent que, pour certaines classes, il est préférable d'utiliser uniquement les BoW, pour d'autres d'utiliser seulement les BoVW et pour la plupart d'entre elles la fusion. Le choix entre les trois méthodes doit être guidé par la connaissance de l'utilisateur, qui peut varier en fonction du contexte. Si les documents sont de bonne qualité (300 dpi ou plus, texte lisible, structure simple, etc.) et contiennent beaucoup de mots, les meilleures performances seront obtenues avec les BoW. Si les documents ont des défauts, des transformation affine géométrique, des écritures manuscrites etc., les meilleures performances seront obtenues en utilisant les BoVW. Enfin, si la base de données est totalement inconnue, s'il est difficile de prédire si l'OCR va bien fonctionner ou si les documents on des contenus très hétérogènes nous conseillons d'utiliser la combinaison des BoW et des BoVW.

Temps d'exécution

En moyenne, l'application de l'OCR sur des documents prend environ 3 secondes par image. L'extraction des points d'intérêt (sur des images redimensionnées à 30 %) prend environ $0,9$ secondes. La phase d'apprentissage se fait hors-ligne. L'apprentissage pour les BoW ne prend que quelques minutes, car il consiste simplement à filtrer et compter des mots. L'étape de l'apprentissage des BoVW prend plus de temps parce que l'algorithme k-means doit être appliqué sur des millions de points d'intérêt, quelques heures sont donc nécessaires (de nombreuses adaptations de k-maeans existent pour le rendre plus rapide). Enfin, la reconnaissance est faite en ligne par les SVM. Il faut environ 20 millisecondes pour prendre une décision.

5. Conclusion et perspectives

Dans cet article, nous avons montré que la méthode des BoVW, qui est habituellement utilisée pour des images naturelles, peut également être appliquée sur des images de documents, fournissant des résultats intéressants quand l'OCR échoue à extraire des informations textuelles.

Même si, de manière général, les BoVW fonctionnent un peu moins bien que les BoW, nous montrons que l'utilisation d'une combinaison des deux techniques (fusion par apprentissage) améliore le rappel et la précision de la classification des images de documents.

De nombreuses perspectives sont à explorer afin d'améliorer la classification multimodale d'images de documents dans un contexte industriel. Par exemple, nous envisageons de tester d'autres techniques de combinaison, d'utiliser différentes bases de données d'apprentissage pour le texte et les caractéristiques de l'image, de choisir le meilleur classifieur pour chaque document au lieu de combiner systématiquement les classifieurs et enfin d'ajouter d'autres informations telles que la mise en page.

Nous travaillons également à la mise en place d'une base de tests qui serait publique. Étant donnée que certaines images contiennent des données personnelles, elles ne peuvent pas être diffusées en l'état. Nous envisageons donc de mettre à disposition, pour chaque image, la vérité terrain qui lui est associée et le résultats du calcul de descripteurs de BOW et BOVW.

Remerciements

Les auteurs tiennent à remercier Jean-Marc Nahon, le directeur de l'informatique de l'entreprise Gestform. Toutes les images sont fournies par Gestform.

6. Bibliographie

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

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

Csurka G., Dance C., Fan L., Willamowski J., Bray C., « Visual categorization with bags of keypoints », *Workshop on statistical learning in computer vision, ECCV*, vol. 1, p. 22, 2004.

Gordo A., Rusinol M., Karatzas D., Bagdanov A. D., « Document Classification and Page Stream Segmentation for Digital Mailroom Applications », *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, IEEE, p. 621-625, 2013.

Gorgevik D., Cakmakov D., « Handwritten digit recognition by combining SVM classifiers », *Computer as a Tool, 2005. EUROCON 2005. The International Conference on*, vol. 2, IEEE, p. 1393-1396, 2005.

Ho T. K., Hull J. J., Srihari S. N., « Decision combination in multiple classifier systems », *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 16, n° 1, p. 66-75, 1994.

Joachims T., *Learning to classify text using support vector machines : Methods, theory and algorithms*, vol. 186, Kluwer Academic Publishers Norwell, MA, USA :, 2002.

Kittler J., Hatef M., Duin R. P., Matas J., « On combining classifiers », *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, n° 3, p. 226-239, 1998.

Liu B., « Sentiment analysis and opinion mining », *Synthesis Lectures on Human Language Technologies*, vol. 5, n° 1, p. 1-167, 2012.

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

Meüller H., Clough P., Deselaers T., Caputo B., *ImageCLEF : Experimental Evaluation in Visual Information Retrieval*, vol. 32, Springer, 2010.

Platt J., « Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods », *Advances in large margin classifiers*, vol. 10, n° 3, p. 61-74, 1999.

Rusinol M., Karatzas D., Bagdanov A. D., Lladós J., « Multipage document retrieval by textual and visual representations », *Pattern Recognition (ICPR), 2012 21st International Conference on*, IEEE, p. 521-524, 2012.

Rusinol M., Lladós J., « Logo spotting by a bag-of-words approach for document categorization », *2009 10th International Conference on Document Analysis and Recognition*, IEEE, p. 111-115, 2009.

Sebastiani F., « Machine learning in automated text categorization », *ACM computing surveys (CSUR)*, vol. 34, n° 1, p. 1-47, 2002.

Shekhar R., Jawahar C., « Word Image Retrieval Using Bag of Visual Words », *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*, IEEE, p. 297-301, 2012.

Snoek C. G., Worring M., « Multimodal video indexing : A review of the state-of-the-art », *Multimedia tools and applications*, vol. 25, n° 1, p. 5-35, 2005.

Song W., Uchida S., Liwicki M., « Look Inside the World of Parts of Handwritten Characters », *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, IEEE, p. 784-788, 2011.

Terrades O. R., Valveny E., Tabbone S., « Optimal classifier fusion in a non-bayesian probabilistic framework. », *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, n° 9, p. 1630-1644, 2009.

Valle E., Cord M., « Advanced Techniques in CBIR : Local Descriptors, Visual Dictionaries and Bags of Features », *Tutorials of the XXII Brazilian Symposium on Computer Graphics and Image Processing*, IEEE, p. 72-78, 2009.

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

Ye G., Liu D., Jhuo I.-H., Chang S.-F., « Robust late fusion with rank minimization », *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, IEEE, p. 3021-3028, 2012.