Extraction des bulles de bandes dessinées [2]

Christophe Rigaud — Jean-Christophe Burie — Jean-Marc Ogier

Laboratoire L3i, Université de La Rochelle
Avenue Michel Crépeau 17042 La Rochelle Cedex 1 - France
{christophe.rigaud, jcburie, jmogier}@univ-lr.fr

RÉSUMÉ. *Les bandes dessinées et les mangas sont l'une des formes les plus populaires de contenu graphique dans le monde et jouent un rôle majeur dans la diffusion de la culture. Aujourd'hui, la numérisation massive permet la lecture page par page mais nous pensons que d'autres usages peuvent émerger. Dans cet article, nous nous concentrons sur l'extraction de phylactères qui est une étape majeure pour permettre l'association texte/graphique dans les images de bandes dessinées. En effet, les bulles de dialogues sont à l'interface entre le texte et les personnages, elles informent le lecteur du ton et de la position des locuteurs. Nous présentons une méthode de segmentation de phylactères génériques et indépendante du texte, basée sur les niveaux de gris, la forme et l'organisation topologique des composantes connexes. La méthode a été évaluée sur les ensembles de données publics eBDtheque et Manga109, les résultats de F-mesure obtenus sont respectivement de 78,24% et 80.04%.*

ABSTRACT. *Comics and manga are one of the most popular and familiar forms of graphic content over the world and play a major role in spreading country's culture. Nowadays, massive digitization allow page-per-page mobile reading but we believe that other usages will be released in the near future. In this paper, we focus on speech balloon segmentation which is a key issue for text/graphic association in scanned and digital-born comic book images. Speech balloons are at the interface between text and comic characters, they inform the reader about speech tone and the position of the speakers. We present a generic and text-independent speech balloon segmentation method based on grey levels, shape and topological organization of the connected-components. The method has been evaluated at pixel-level on two public datasets (eBDtheque and Manga109) and the F-measure results are 78.24% and 80.04% respectively.*

MOTS-CLÉS : *reconnaissance graphique, bulle de dialogue, analyse d'images de BD.*

KEYWORDS: *graphic recognition, speech balloon, comics image analysis, manga image analysis.*

2. Traduction étendue d'un article publié en anglais à GREC 2015 (Rigaud *et al.*, 2015a)

1. Introduction

Les ventes de bandes dessinées numériques atteignent maintenant 10% du marché de la bande dessinée et ont doublé au cours des cinq dernières années[1]. Cette nouvelle manière de lire permet de nouveaux usages grâce à la richesse des dessins et au développement récent d'outils de lecture mobile. En dehors de la réorganisation automatique des cases en fonction de la taille de l'écran, il y a peu de travaux explorant d'autres modes de lecture.

Dans cet article, nous nous focalisons sur la détection des bulles au niveau pixel afin d'extraire leurs position et forme. Ces deux informations sont essentielles pour la compréhension automatique des bandes dessinées, en particulier pour la classification de bulles (Rigaud *et al.*, 2014) et l'association des personnages aux bulles (Rigaud *et al.*, 2015c). Cette dernière n'est pas explicitement dessinée par l'auteur mais implicitement comprise par le lecteur en fonction de la position des éléments dans les planches. Les bulles sont placées de manière à aider le lecteur à les associer aux personnages et à suivre l'histoire. Les cases, les bulles et les personnages sont les trois informations nécessaires pour associer les bulles et les personnages au sein des cases. L'extraction de cases est la tâche la plus simple dans l'analyse d'images de bandes dessinées et plusieurs études proposent des méthodes capables de dépasser 80% de rappel et précision (Stommel *et al.*, 2012 ; Li *et al.*, 2015). L'extraction des personnages est à ces débuts et la connaissance de la position des bulles ainsi que de leurs queues facilitent l'extraction de tels objets graphiques (Rigaud *et al.*, 2015c). L'extraction des phylactères au niveau pixel (et non pas au niveau boîtes englobantes) est essentiel pour une analyse ultérieure de leurs contour et queues (Figure 1).

Nous proposons une approche indépendante du texte pour l'extraction de bulles de dialogue dites fermées (bulles avec un contour entièrement fermé, les plus répandus). Nous basons notre approche sur l'hypothèse que les bulles de dialogue sont des régions qui contiennent des éléments alignés et contrastés (propriété du texte). Dans ce qui suit, nous allons utiliser l'acronyme *BD* pour désigner tous les types de bandes dessinées y compris les Mangas (bandes dessinées japonaises).

Les bulles sont des éléments clés dans la bande dessinée, elles contiennent la plupart de l'information textuelle et vont de paire avec les personnages (les locuteurs). Peu de travaux sur l'extraction de bulles ont été publiés jusqu'à aujourd'hui. Ils portent principalement sur les bulles fermées. Ces travaux sont basés sur l'analyse de composantes connexes. Arai a proposé une méthode de détection de zones blanches à l'aide de quatre règles de filtrage propres aux images de mangas (Arai et Tolle, 2011). Ces règles sont basées sur la taille, le nombre de pixels blancs, la présence d'espaces blancs verticaux et le ratio hauteur/largeur. Plus récemment, une proposition d'utilisation de l'espace colorimétrique HSV pour faire une première sélection des régions claires étant ensuite considérées comme étant des bulles avec un rapport entre la zone de texte et leurs boîtes englobantes supérieure à 60% (Ho *et al.*, 2012). Récemment, Liu

1. Livre blanc de Milton Griepp, Conférence ICv2 2014

(a) Bulle rectangulaire sur fond grisé avec une queue

(b) Bulle ondulée sur fond jaunit avec une queue

(c) Bulle mixte (ovale/rectangle) sur fond clair avec une queue

(d) Bulle mixe (ondulée/rectangle) sur fond jaunit sans queue

Figure 1. *Détection de bulles au niveau pixel sur différents type de bulles et arrière plan (ligne rouge).*

et al. ont proposé d'apprendre les caractéristiques des bulles contenant du texte afin de reconnaître les bulles fermées avec un fond blanc à l'aide de classificateurs (Liu *et al.*, 2015a).

L'extraction de bulles non fermées (bulles avec des contours partiellement dessinés ou absents) a été très peu étudiée, une première approche a été proposée en étendant un modèle de contour actif autour des zones de texte (Rigaud *et al.*, 2013). Liu *et al.* ont proposé d'appliquer une croissance de région à partir des zones de texte puis de séparer en deux la région résultante au niveau de la zone la plus étroite (embouchure de la queue), dans le cas où elle se serait beaucoup étendue et éloignée de la zone de texte initiale (Liu *et al.*, 2015b). Les deux approches requièrent la position des zones de texte en entrée.

La section suivante présente la méthode de segmentation de bulles proposée. La Section 3 détaille les expérimentations que nous avons mené. Enfin, les Sections 4 et 5 discutent les résultats et tirent les conclusions de ce travail.

2. Extraction de bulles

Parmi les cinq approches présentées en Section 1, la première a été développée spécialement pour les images de mangas et montre des difficultés à traiter d'autres types de BD présentant des caractéristiques différentes. Tout d'abord, l'extraction des composantes connexes (CC) nécessite une image binaire qui est obtenue à l'aide d'un seuil global dans cette approche. Par conséquent, cela contraint son application aux images avec une couleur de fond très claire (tendant vers le blanc). Par ailleurs, la sélection des régions candidates est réalisée en utilisant plusieurs heuristiques qui ne sont pas validées expérimentalement et qui sont spécifiques aux mangas. La méthode proposée par Ho (Ho *et al.*, 2012) peut s'avérer très efficace pour un type de bandes dessinées en particulier, mais l'ensemble des paramètres utilisés rend sa généralisa-

tion à tous les types de bandes dessinées très difficile (e.g. pourcentage minimum de texte à l'intérieur des bulles). La méthode proposée par Liu *et al.* obtient des résultats satisfaisants mais elle nécessite également des bulles avec un fond blanc (seuillage binaire fixe). Cette méthode dépend aussi du jeu de donnée utilisé pour l'apprentissage (supervisé). Les deux travaux concernant l'extraction des bulles ouvertes requièrent la positions *a priori* du texte (donnée d'entrée) ce qui est une contrainte forte en raison du problème de propagation d'erreur (provenant de l'extraction de texte préalable). Néanmoins, cette méthode a l'avantage d'extraire les bulles ouvertes et fermées simultanément (Rigaud *et al.*, 2013 ; Liu *et al.*, 2015b).

Nous proposons de dépasser ces limitations en utilisant une méthode sans apprentissage sur la base d'un seuillage adaptatif puis de d'extraire et d'analyser des composantes connexes. Comme pour l'analyse d'images de documents en général, cette méthode limite la division intempestive de traits continus (Figure 2). Après avoir extrait toutes les CC, nous sélectionnons uniquement celles qui présentent des particularités de contraste, de topologie et de forme, indépendamment de leur taille et de la nature de leur contenu (e.g. symboles, lettres) puis on calcule un indice de confiance qui leur est associé et qui est utilisé pour la classification finale en bulle/non bulle.

2.1. *Sélection adaptative du seuil*

Au cours de la création de bandes dessinées ou de mangas, les contours des bulles sont traditionnellement dessinés à l'aide d'un trait noir, puis rempli avec du texte (Cyb, 2006). Nous nous basons sur ces deux informations qui sont intrinsèques au processus de création et qui sont donc des caractéristiques des bulles de dialogues.

Le contour des bulles est dessiné sous la forme d'un trait continu représentant une succession de lignes droites ou courbes. Parfois, ce trait est dégradé en raison des techniques de numérisation des images ou de la compression. Une segmentation parfaite du contour se traduirait par une seule composante connexe par contour de bulle (Figure 2). Néanmoins, les régions ayant un fond complexe compliquent cette étape. Il existe plusieurs méthodes de sélection de seuil adaptatif dans la littérature (Lamiroy et Ogier, 2014), cependant, les phylactères étant des régions très contrastées, la séparation entre les traits sombres et l'arrière-plan (généralement blanc) en est facilitée. Les principales difficultés sont la forme et la taille de la fenêtre glissante qui est utilisée pour déterminer si un pixel appartient à l'arrière-plan ou au premier plan. N'ayant pas *a priori* d'informations sur la localisation, la forme et la taille des bulles, nous définissons une fenêtre carrée de taille $blockSize$ relative à la taille de l'image. Nous définissons la valeur de seuil $T(x, y)$ telle que la moyenne de la région de taille $blockSize * blockSize$ centrée sur (x, y). Le pixel correspondant à la position (x, y) est considéré comme faisant partie de premier plan si sa valeur de niveau de gris est supérieure à T, sinon il fait partie de l'arrière plan. La Figure 3 montre les résultats du seuillage adaptatif de sous quadrants carrés de taille $blockSize$ appliqué à des BD de nature et résolution différentes.

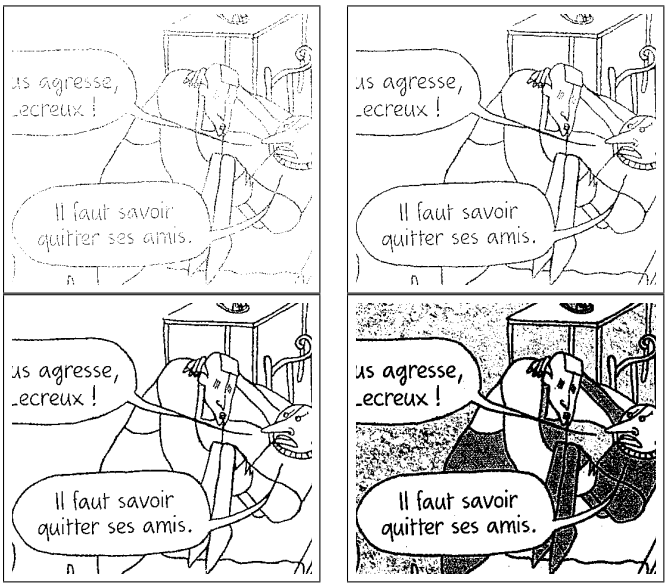
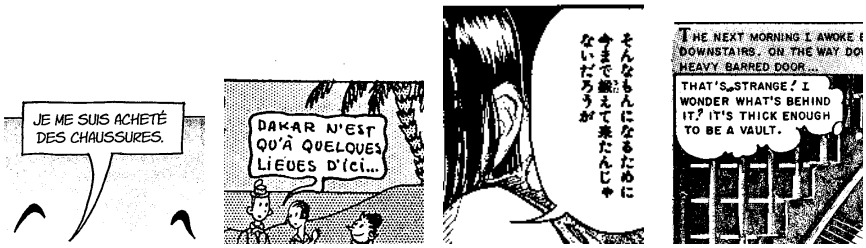


Figure 2. *Résultats de seuillages globaux appliqués à une image en niveau de gris sur 8 bits, du plus faible (en haut à gauche) au plus élevé (en bas à droite) avec des seuils de 50, 100, 150, 200. Nous observons que les traits noirs sont discontinus pour les seuil bas et le fond commence à apparaître sous la forme de bruit poivre/sel en raison de la texture du papier à une valeur de seuil haute. Le seuil optimal est 150 dans cet exemple car la continuité des traits est préservée et peu d'éléments superflus apparaissent. Crédits image : (Roudier, 2011).*

Après le seuillage de l'image, nous obtenons une image binaire à partir de laquelle nous allons extraire puis analyser les relations entre les composantes connexes blanches et noires en utilisant un algorithme d'étiquetage de composantes connexes proposé par Suzuki (Suzuki *et al.*, 1985). Les CC sont ensuite divisées en deux ensembles selon leur niveau de gris afin de faciliter les traitements ultérieurs (séparation contour/contenu). Les ensembles de CC blanches et noires sont appelés W et B respectivement. Notons que d'autres techniques d'extraction de régions pourraient être utilisées (e.g. MSER (Donoser et Bischof, 2006)) mais un traitement supplémentaire devrait être ajouté afin de retrouver les relations topologiques entre les CC.

2.2. *Sélection des CC candidates*

Les bulles peuvent être considérées comme des conteneurs d'éléments principalement graphiques ce qui implique que le texte est inclus à l'intérieur des bulles. D'un point de vue topologique, la région englobant le contenu d'une bulle est le fond de



(a) BD française récente (format A4 numérisée à 300DPI) (b) BD française ancienne (format A4 numérisée à 300DPI) (c) Manga récent (format A5 numérisé à 150DPI) (d) BD américaine de l'age d'or (format A4 numérisée à 300DPI)

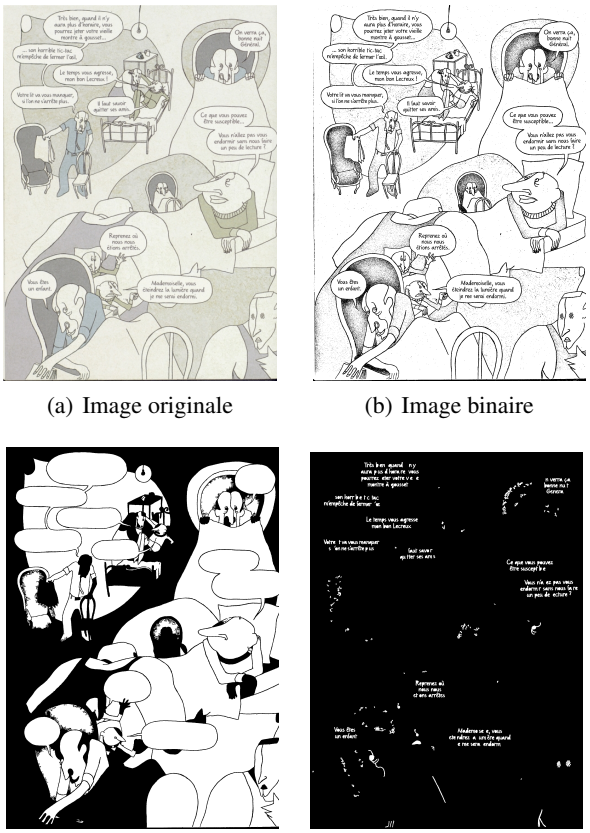
Figure 3. *Résultats après seuillage de différents types de BD scannées à différentes résolutions. Les images originales sont en Figure 1.*

cette même bulle (e.g. région blanche entourée par un contour noir). Notons que le fond des bulles est supposé faire partie de l'ensemble W car le fond des bulles est habituellement plus clair que leur contenu qui lui fait partie de l'ensemble B. Nous proposons de nous appuyer sur des caractéristiques de colorimétrique et topologique pour sélectionner parmi l'ensemble W les CC (parents) contenant d'autres CC (enfants) qui font partie de l'ensemble B. Nous appelons ce sous-ensemble de parents les "CC candidates" (Figure 4). Notons que la plus grande CC blanche, correspondant généralement à l'arrière plan de l'image, est ignorée ainsi que les très petites régions habituellement issues d'erreurs de seuillage de la texture du papier ou de bruits liés à la compression de l'image ($< 0,5\%$ de la taille de l'image).

Dans la sous-section suivante, la forme et l'organisation spatiale de chaque CC candidate sont analysées de manière à déterminer si elles contiennent ou non des éléments semblable à du textes (CC alignées).

2.3. *Analyse des CC candidates*

Nous proposons d'analyser l'organisation du contenu de chaque CC candidates afin de déterminer si elles contiennent des éléments similaires à du texte (sous l'hypothèse que les bulles de dialogues contiennent du texte). Le texte parlé présente plusieurs caractéristiques, certaines propres au texte et d'autres à la bande dessinée. Il présente aussi certaines caractéristiques qui sont indépendantes de la langue. C'est le cas pour l'alignement et l'espacement des glyphes (plus petite portion d'une lettre ou d'un symbole) avec un contraste important, la régularité du trait, la couleur ou la taille similaire (Bigorda et Karatzas, 2014). Lorsque le texte est utilisé dans des dialogues, il est la plupart du temps aligné et centré à l'intérieur d'une région en forme de bulle. Cependant, l'espacement des glyphes et la largeur du trait ne sont pas toujours stables car le texte est souvent manuscrit dans les BD. La difficulté dans l'analyse d'image de



(a) Image originale (b) Image binaire

(c) Ensemble des parents (W) (d) Ensemble des enfants (B)

Figure 4. *Ensemble des CC candidate (W) et leurs contenu (B) for une image donnée. Les CC correspondantes sont représentés en blanc pour chaque ensemble. Crédits image : (Roudier, 2011).*

bandes dessinées est la quantité importante de graphiques qui sont également constitués d'éléments alignés, ce qui perturbent le processus de reconnaissance des bulles parmi l'ensemble des éléments dessinés (e.g. des tuiles, de l'herbe, des cheveux, les yeux).

Dans l'approche ci-dessous, nous combinons les caractéristiques d'alignement et de forme pour calculer un indice de confiance pour chaque CC candidates. L'indice de confiance est utilisé pour la décision finale pour déterminer si chaque CC candidate correspond effectivement à une bulle ou non (Section 3).

(a) 0/18 CC alignées (b) 7/18 CC alignées, 1 ligne trouvée (c) 14/18 CC alignées, 2 lignes trouvées (d) 16/18 CC alignées, 3 lignes trouvées

Figure 5. *Processus d'association par alignement des CC enfants. Cette méthode s'arrête automatiquement lorsqu'il y reste moins de CC non alignées que de lignes trouvées (e.g. Figure 5(d), trois lignes ont été construites et seulement deux CC enfants sont restantes).*

2.3.1. *Alignement du contenu*

Les CC enfants sont supposées être alignées horizontalement ou verticalement selon la langue (ex : vertical pour le japonais et horizontal pour les langues latines). Ceci est une caractéristique du texte de dialogue dans la bande dessinée mais aussi pour le texte en général. Nous proposons de "scanner" le contenu de chaque CC candidate et de calculer le pourcentage d'enfants alignés appelé alignement inter-enfant ($cAlign$). La direction de "scannage" (de haut en bas ou de gauche à droite) est définie manuellement ici, elle pourrait être défini automatiquement en appliquant systématiquement les deux directions séparément puis en sélectionnant celle qui produit les meilleurs résultats (Section 2.3.3).

Nous calculons le pourcentage de CC enfants alignées dans un ordre spécifique, du plus long au plus court afin de trouver les lignes les plus longues en premiers (les plus représentatives de d'information textuelle). Deux CC enfants sont considérées comme alignées horizontalement si le centre de la première CC se situe entre la valeur verticale minimale et maximale de la seconde. Par exemple, pour deux enfants A et B et leur centres respectifs (cx, cy), B est aligné horizontalement avec A si $A.ymin < B.cy < A.ymax$. De la même manière, l'alignement vertical est calculé en remplaçant y par x. Notons que trois enfants au minimum sont nécessaires pour calculer un alignement pertinent. Le processus s'arrête automatiquement quand il y a moins d'enfants non alignés restant que le nombre de lignes déjà trouvées. Ce technique permet d'ignorer les enfants non-alignés correspondant aux ponctuations, aux accents, aux composantes connexes discontinues, etc. (Figure 5).

2.3.2. *Analyse de forme*

La forme des bulles de dialogue est semblable à celle d'une bulle (de savon) qui contient des éléments (principalement du texte). Son contour est généralement maté-

rialisé par un trait qui a quelques irrégularités en fonction de l'émotion que l'auteur souhaite retranscrire chez le lecteur. Notons que ces irrégularités créent des défauts de convexité dont les deux plus importants se trouvent dans la région de la queue (Rigaud *et al.*, 2015b). Nous proposons de mesurer la convexité globale du contour afin de trouver son degré de similitude avec une forme totalement convexe (e.g. cercle, ovale, rectangle, etc.). Plusieurs mesures de convexité existent dans la littérature (Chalmeta *et al.*, 2013), certaines sont basées sur la surface et d'autres sur le périmètre de la forme. Nous avons opté pour une approche fondée sur l'analyse du périmètre car il est beaucoup plus influencé par les petites variations de contours que la forme générale. Nous définissons la mesure de la convexité ($cShape$) comme étant le rapport entre le périmètre Euclidien de l'enveloppe convexe de la forme S et le périmètre Euclidien de la forme elle-même (Équation 1).

$$cShape = \frac{arcLength(hull(S))}{arcLength(S)} \quad [1]$$

Notons que la mesure de convexité est égale à 100% pour une une forme parfaite comme les rectangles, carrés, ovale, cercles, etc. Cela pourrait être le cas pour les phylactères sans queue, mais pas pour les autres car la région de la queue est habituellement non convexe de manière à clairement indiquer la position du personnage. Cependant, le périmètre de la région de la queue ne représente qu'une petite partie de l'ensemble du périmètre d'une bulle et donc a un impact mineur sur la mesure proposée dans la majorité des cas. De plus, les régions des queues peuvent être détectées et supprimées en utilisant une approche de détection de queue (Rigaud *et al.*, 2015b) mais nous avons préféré éviter toute propagation d'erreur et donc ne pas l'utiliser ici.

2.3.3. *Indice de confiance*

L'indice de confiance global C est calculé pour chaque CC candidate, à partir des indices d'alignement des CC enfants $vAlign$ et de forme $cShape$ selon la Formule 2.

$$C = cAlign * \alpha + cShape * \beta \quad [2]$$

où α et β sont deux variables de pondération dont la valeur a été défini expérimentalement (Section 3.2). C est exprimé en pourcentage.

3. Expérimentations

Dans cette section nous évaluons la méthode proposée de segmentation de bulles de dialogue en utilisant deux jeux de données publics et comparons nos résultats à d'autres approches de la littérature.

3.1. *Jeux de données*

Nous évaluons la méthode proposée en utilisant les jeux de données publics eBDtheque (Guérin *et al.*, 2013) et Manga109 (Matsui *et al.*, 2015) afin de montrer la robustesse de la méthode proposée pour les bandes dessinées comportant du texte en Français, Anglais et Japonais (indépendance aux types d'alphabets ou syllabaires).

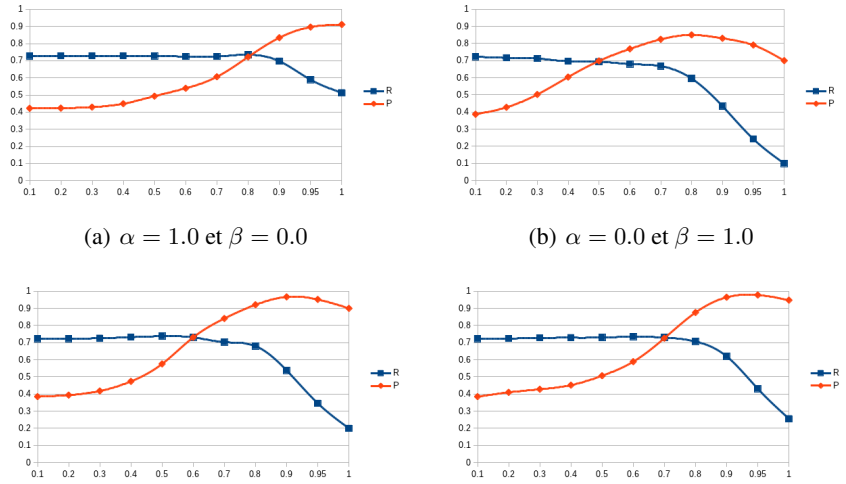
Le jeu de données eBDtheque a été conçu pour être aussi représentatif que possible de la diversité des types de bandes dessinées, il comprend donc quelques pages de divers albums. Il est composé d'une centaine d'images contenant 850 cases, 1550 instances de personnages, 1092 bulles (84,5% sont fermées) et 4691 lignes de texte au total. Il contient des images numérisées à partir d'album de BD françaises (46%), de web comics français (37%) de différents formats et résolutions d'images, de comics américains (11%) et d'illustrations inédites de mangas (6%). En plus de la diversité de styles, des formats et de résolutions d'images, il y a aussi des différences dans les techniques de conception et d'impression puisque 29% des images ont été publiées avant 1953 et 71% après 2000. Ce jeu de données fournit également la position des bulles au pixel près sous forme de vérité terrain.

Le jeu de données Manga109 est une sélection de 109 albums de mangas publiés entre 1970 et 2010 (21142 images au total). Ces albums ont été sélectionnés à partir de l'archive "Manga Library Z" géré par J-comi[2]. "*The Manga109 dataset covers various kinds of categories, including humor, battle, romantic comedy, animal, science fiction, sports, historical drama, fantasy, love, romance, suspense, horror, and four-frame cartoons*" (Matsui *et al.*, 2015). La transcription du texte de seulement quatre des albums est disponible dans la vérité terrain associée. Néanmoins, les auteurs fournissent l'outil de création de la vérité terrain. Cet ensemble de données étant assez conséquent, la création de la vérité terrain au niveau des pixels demanderait beaucoup de temps. Nous avons préféré sélectionner quelques albums uniquement et compter manuellement le nombre de bulles correctes/incorrectes/manquées pour chaque image. Nous avons choisi au hasard un sous-ensemble de trois albums qui représentent 408 images et 3242 bulles au total ("Momoyama Haikagura", "Tetsu San" et "Ultra Eleven").

3.2. *Validation de l'indice de confiance*

Les deux mesures présentées en Section 2.3 ont été évaluées séparément puis combinées de manière à définir les valeurs des variables de pondération α et β de la Formule 2. Quelques résultats obtenus avec le jeu de donnée eBDtheque sont illustrés Figure 6. Les meilleures performances ont été obtenues pour $\alpha = 0.75$ et $\beta = 0.25$.

2. http ://www.j-comi.jp/



(a) $\alpha = 1.0$ et $\beta = 0.0$ (b) $\alpha = 0.0$ et $\beta = 1.0$

(c) $\alpha = 0.5$ et $\beta = 0.5$ (d) $\alpha = 0.75$ et $\beta = 0.25$

Figure 6. *Performances de rappel R et de précision P obtenues pour chaque niveau de confiance C et pour différentes valeurs de α et β. Pour chaque mesure, seules les CC candidates avec un indice de confiance supérieur ou égal à C, variant de 10 à 100% sur l'axe des abscisses, sont considérées comme valides (vrais positifs).*

3.3. *Évaluation de performance*

Nous avons évalué la performance de la méthode proposée aux niveaux pixels et boîtes englobantes pour le jeu de données eBDtheque afin de présenter des résultats à la fois précis et comparables avec d'autres méthodes de la littérature fournissant que des résultats au niveau des boîtes englobantes par exemple. Cependant, les résultats sur le sous-ensemble du jeu Manga109 ont seulement été évalués visuellement au niveau objet (les bulles) en raison de l'absence de vérité terrain. Ces derniers ont été réalisés pour les CC candidates ayant un indice de confiance $C \geq 80\%$ au minimum afin d'évaluer la performance des vingt meilleurs pourcent fournit par la méthode proposée. En ce qui concerne la sélection du seuil adaptatif, la valeur $blockSize$ a été définie comme un carré de surface égale à 1,3% de l'aire de l'image globale à partir d'une validation expérimentale faite sur le jeu de données eBDtheque (Guérin *et al.*, 2013).

$$R = \frac{TP}{TP + FN} \quad [3]$$

$$P = \frac{TP}{TP + FP} \quad [4]$$

(a) Considéré FP au niveau boîte englobante ($a_0 < 0.5$)

(b) Considéré TP au niveau pixel ($>$ 90% pixels en commun)

Figure 7. *Différence de résultats obtenus pour une même bulle avec deux niveaux d'évaluations (niveaux boîte englobante et pixel). Les polygones rouges et bleus correspondent respectivement aux régions détectées et et à celles de la vérité terrain.*

Nous allons maintenant détailler l'évaluation réalisée au niveau pixel. A ce niveau, chaque pixel de chaque CC candidates est considéré comme vrai positif TP si il correspond à un pixel d'une bulle appartenant à la vérité terrain, sinon faux positif FP. Les mesures de TP, FP et faux négatifs FN (pixels manqués) sont utilisés pour calculer le taux de rappel R et de précision P de chacune des méthodes selon les Formules 3 et 4. Nous avons également calculé la F-mesure F pour chaque résultat.

Concernant l'évaluation au niveau des boîtes englobantes, nous utilisons la même métrique que le challenge PASCAL VOC pour les objets visuels (Everingham *et al.*, 2010). La correspondance entre les régions détectées et les régions de la vérité terrain est effectuée selon leur recouvrement mutuel. Pour qu'une région détectée soit considérée comme vraie positive (TP), le taux de recouvrement a_0 entre sa boîte englobante B_p et son équivalent dans la vérité terrain B_{GT} doit dépasser 0,5 (Formule 5).

$$a_0 = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \quad [5]$$

Selon le challenge PASCAL VOC, les objets détectés sont considérés comme vrais positifs TP si $a_0 > 0,5$ ou faux positif FP autrement. Les bulles manquées sont comptabilisées comme faux négatifs (FN). Notons que l'évaluation au niveau des boîtes englobantes est moins précise que l'évaluation réalisée au niveau des pixels, en particulier pour les phylactères ayant une longue queue (Figure 7).

3.4. *Analyse des résultats*

Pour le premier jeu de données (eBDtheque), nous avons comparé les résultats de la méthode proposée avec d'autres méthodes de la littérature aux niveaux pixels et

boîtes englobantes. Les résultats sont présentés dans le Tableau 1. Notons que Liu *et al.* (Liu *et al.*, 2015b) ont indiqué des résultats au niveau boîte englobante seulement et en utilisant un ratio de recouvrement légèrement plus contraignant ($a_0 > 0.6$).

Tableau 1. *Performance moyenne de segmentation des bulles en pourcentage au niveau pixel et boîte englobante (BdB).*

	Niveau pixel			Niveau BdB		
Méthodes	R	P	F_1	R	P	F_1
Arai (Arai et Tolle, 2011)	18.70	23.14	20.69	13.40	11.76	12.53
Ho (Ho *et al.*, 2012)	14.78	32.37	20.30	13.96	24.76	17.84
Rigaud (Rigaud *et al.*, 2013)	69.81	32.83	44.66	52.68	44.17	48.05
Liu (Liu *et al.*, 2015b)	–	–	–	**80.10**	75.60	**77.80**
Proposed	**70.71**	**87.62**	**78.24**	72.21	**83.31**	77.36

Au niveau pixel, la méthode proposée fournit les meilleurs résultats par rapport aux autres méthodes de la littérature, y compris pour une méthode qui utilise un scénario simplifié nécessitant la position *a priori* du texte (Rigaud *et al.*, 2013). Le rappel de 70,71% a été mesuré sur l'ensemble des bulles contenues dans le jeu eBDtheque dont 15,5% d'entre elles sont ouvertes (non détectables par l'approche proposée). Les 13,79% d'erreur restants sont essentiellement dû à de petites bulles contenant très peu d'informations ou des dessins (Figure 8(a)). Les régions qui mettent en échec l'approche proposée (baisse de la précision) sont souvent constituées de texte illustratif ou d'éléments textuels comme le montrent les Figures 8(d) et 8(e)).

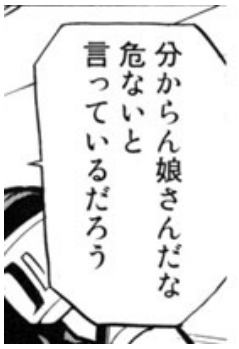
En ce qui concerne l'expérimentation sur le sous-ensemble du jeu de données Manga109, nous obtenons des performances moyennes de taux de rappel, précision et F-mesure de 72,24%, 89,71% et 80.04% respectivement. La performance globale est similaire à celle obtenue avec le jeu eBDtheque, cela confirme que la méthode proposée est insensible aux styles de bandes dessinées (Européen, Américain, Asiatique, etc.). Des limites semblables à celles du jeu eBDtheque ont été observés. Les baisses de rappel et de précision sont principalement dues aux bulles ouvertes (Figure 8(b)) et au texte illustratif dans des régions très similaires aux bulles (Figure 8(f)).

Concernant l'évaluation au niveau boîtes englobantes, l'approche proposée excède les méthodes de l'état de l'art en terme de précision et de F-mesure, mais pas en rappel. Notons que la méthode proposée a l'avantage de ne pas nécessiter la position du texte en entrée contrairement à la méthode (Liu *et al.*, 2015b).

Les résultats détaillés sont disponibles sur GitHub[3].

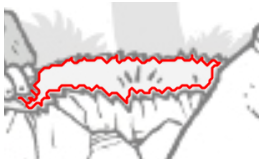
3. `https://github.com/crigaud/publication/tree/master/2016/LNCS/text-independent_speech_balloon_segmentation_for_comics_and_manga`

(a) Moins de trois éléments sont alignés dans les bulles

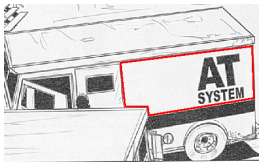


(b) Bulles ouverte non détectable avec la méthode proposée

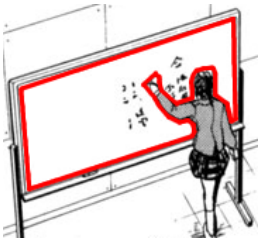
(c) Bulle sans bordures non détectable par la méthode présentée



(d) L'herbe est alignée dans une région similaire à ne bulle



(e) Illustration contenant du texte (bâche de camion)



(f) Illustration contenant du texte (tableau blanc)

Figure 8. *Exemples de cas de mise en échec de l'approche proposée qui engendre une baisse du taux de rappel pour la première ligne (FN) et du taux de précision pour la seconde ligne (FP). Le contour des bulles détectées est représenté par un trait rouge dans la seconde ligne.*

4. Discussion

La méthode proposée utilise un seuillage adaptatif simple et efficace pour notre périmètre d'étude, car les bulles de dialogues sont des régions très contrastées et donc facile à seuiller lorsqu'elles sont pris isolément du reste de l'image (seuillage local). Notons que l'approche proposée fait l'hypothèse que le fond des phylactères est homogène et plus clair que leurs contenus. Si cela n'est pas le cas, une inversion des couleurs de l'image doit être appliquée en amont. La sélection des composantes connexes candidates est basée sur l'analyse de leur contenus qui, parfois, est constitué de graphiques ayant une organisation similaire à du texte (e.g. alignés ou de même taille). L'approche présentée extrait des régions correspondantes aux fonds des bulles (régions blanches), cependant, certaines approches requièrent plus l trait de contour des bulles (e.g. calcul précis de la direction de la queue). Dans ce cas, les résultats pro-

posés devront être post-traités afin d'extraire le bord extérieur du contour des bulles comme l'a proposé (Liu *et al.*, 2015a). Le travail présenté se focalise sur l'analyse de bandes dessinées mais il peut être étendu à d'autres types d'images ayant des relations fortes entre le texte et les formes graphiques, tel que les dessins techniques, les plaques d'immatriculation et les panneaux de signalisation routière).

5. Conclusions

Cet article présente une approche de segmentation de bulles de dialogue dans les images de bandes dessinées et de mangas. La méthode proposée combine des informations colorimétriques, de forme et de relations topologiques des composantes connexes pour segmenter les bulles au niveau pixel. Nous avons également proposé un indice de confiance de la segmentation, basé sur le l'alignement du contenu et la forme des phylactères. Cet indice peut être utilisé pour d'autres applications (e.g. panneaux de signalisation routière). L'approche proposée a été testée sur de nombreux types de bandes dessinées. Ses performances sont très satisfaisantes sur les deux jeux de données publics, à savoir eBDtheque (Guérin *et al.*, 2013) et Manga109 (Matsui *et al.*, 2015). À l'avenir, nous inclurons l'extraction des bulles ouvertes à cette méthode.

Remerciements

Ce travail a été soutenu par l'Université de La Rochelle (France), la ville de La Rochelle et le Programme d'Investissements d'Avenir iiBD. Nous sommes très reconnaissants envers tous les auteurs et éditeurs des jeux de données eBDtheque et Manga109 pour nous avoir permis d'utiliser leurs œuvres.

6. Bibliographie

Arai K., Tolle H., « Method for Real Time Text Extraction of Digital Manga Comic », *International Journal of Image Processing (IJIP)*, vol. 4, n° 6, p. 669-676, 2011.

Bigorda L. G., Karatzas D., « A Fast Hierarchical Method for Multi-script and Arbitrary Oriented Scene Text Extraction », *CoRR*, 2014.

Chalmeta R., Hurtado F., Sacristán V., Saumell M., « Measuring regularity of convex polygons », *Computer-Aided Design*, vol. 45, n° 2, p. 93 - 104, 2013. Solid and Physical Modeling 2012.

Cyb, *Making Comics : Storytelling Secrets of Comics, Manga and Graphic Novels*, William Morrow Paperbacks, p. 128-153, 2006.

Donoser M., Bischof H., « Efficient maximally stable extremal region (MSER) tracking », *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, IEEE, p. 553-560, 2006.

Everingham M., Van Gool L., Williams C. K., Winn J., Zisserman A., « The pascal visual object classes (voc) challenge », *International journal of computer vision*, vol. 88, n° 2, p. 303-338, 2010.

Guérin C., Rigaud C., Mercier A., al., « eBDtheque : a representative database of comics », *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*, Washington DC, p. 1145-1149, 2013.

Ho A. K. N., Burie J.-C., Ogier J.-M., « Panel and Speech Balloon Extraction from Comic Books », *10th IAPR International Workshop on Document Analysis Systems*p. 424-428, March, 2012.

Lamiroy B., Ogier J.-M., « Analysis and Interpretation of Graphical Documents », *in* D. Doermann, K. Tombre (eds), *Handbook of Document Image Processing and Recognition*, Springer, 2014.

Li L., Wang Y., Suen C. Y., Tang Z., Liu D., « A tree conditional random field model for panel detection in comic images », *Pattern Recognition*, vol. 48, n° 7, p. 2129-2140, 2015.

Liu X., Li C., Zhu H., Wong T.-T., Xu X., « Text-aware balloon extraction from manga », *The Visual Computer*p. 1-11, 2015a.

Liu X., Wang Y., Tang Z., « A clump splitting based method to localize speech balloons in comics », *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR).*, IEEE, p. 901-906, 2015b.

Matsui Y., Ito K., Aramaki Y., Yamasaki T., Aizawa K., « Sketch-based Manga Retrieval using Manga109 Dataset », *CoRR*, 2015.

Rigaud C., Burie J.-C., Ogier J.-M., « Text-independent speech balloon segmentation for comics and manga », *Proceedings of the 11th IAPR International Workshop on Graphics Recognition (GREC)*, Loria – Laboratoire Lorrain de Recherche en Informatique et ses Applications (UMR 7503), p. 17-20, 2015a.

Rigaud C., Guérin C., Karatzas D., Burie J.-C., Ogier J.-M., « Knowledge-driven understanding of images in comic books », *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 18, n° 3, p. 199-221, 2015b.

Rigaud C., Karatzas D., Burie J.-C., Ogier J.-M., « Adaptive Contour Classification of Comics Speech Balloons », *in* B. Lamiroy, J.-M. Ogier (eds), *Graphics Recognition. Current Trends and Challenges*, vol. 8746 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, p. 53-62, 2014.

Rigaud C., Karatzas D., Van de Weijer J., Burie J.-C., Ogier J.-M., « An active contour model for speech balloon detection in comics », *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, p. 1240-1244, 2013.

Rigaud C., Le Thanh N., Burie J.-C., Ogier J.-M., Iwata M., Imazu E., Koichi K., « Speech balloon and speaker association for comics and manga understanding », *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR).*, IEEE, p. 351-356, 2015c.

Roudier N., *Les terres creusées*, vol. Acte sur BD, Actes Sud, 2011.

Stommel M., Merhej L. I., Müller M. G., « Segmentation-free detection of comic panels », *Computer Vision and Graphics*, Springer, p. 633-640, 2012.

Suzuki S. *et al.*, « Topological structural analysis of digitized binary images by border following », *Computer Vision, Graphics, and Image Processing*, vol. 30, n° 1, p. 32-46, 1985.