UNIVERSITÉ DE LA ROCHELLE

ÉCOLE DOCTORALE S2IM

Laboratoire Informatique, Image et Interaction (L3i)

THÈSE présentée par :

Christophe RIGAUD

soutenue le : **11 décembre 2014**
pour obtenir le grade de : **Docteur de l'université de La Rochelle**
Discipline : **informatique et applications**

Segmentation et indexation d'objets complexes dans les images de bandes déssinées

---

JURY :

Bart LAMIROY	Maître de conférences, HDR, Université de Lorraine (France), Examinateur, Président du jury
Simone MARINAI	Professeur associé, Université de Forence (Italie), Rapporteur
Apostolos ANTONACOPOULOS	Professeur associé, Université de Salford (Grande-Bretagne), Rapporteur
Jean-Philippe DOMENGER	Professeur, Université de Bordeaux (France), Examinateur
Nicholas JOURNET	Maître de conférences, Université de Bordeaux (France), Examinateur
Jean-Christophe BURIE	Professeur, Université de La Rochelle (France), Directeur de thèse
Dimosthenis KARATZAS	Professeur associé, Université Autonome de Barcelone (Espagne), Encadrant de thèse
Jean-Marc OGIER	Professeur, Université de La Rochelle (France), Encadrant de thèse

Segmentation and indexation of complex objects in comic book images

A dissertation submitted by **Christophe Rigaud** at Universitat Autònoma de Barcelona to fulfil the degree of **Doctor of Philosophy**.

Bellaterra, September 29, 2014

Director	**Prof. Dr. Jean-Christophe Burie** Laboratoire Informatique, Image et Interaction Université de La Rochelle (France)
Co-Directors	**Dr. Dimosthenis Karatzas** Centre de Visió per Computador Universitat Autònoma de Barcelona (Spain) **Prof. Dr. Jean-Marc Ogier** Laboratoire Informatique, Image et Interaction Université de La Rochelle (France)
Thesis committee	**Dr. Bart Lamiroy** Laboratoire Lorrain de Recherche en Informatique et ses Applications Université de Lorraine (France) **Prof. Dr. Jean-Philippe Domenger** Laboratoire Bordelais de Recherche en Informatique Université de Bordeaux (France) **Dr. Nicholas Journet** Laboratoire Bordelais de Recherche en Informatique Université de Bordeaux (France)
European evaluators	**Dr. Simone Marinai** Dipartimento di Ingegneria dell'Informazione Università degli Studi di Firenze (Italy) **Dr. Apostolos Antonacopoulos** The School of Computing, Science & Engineering University of Salford (United Kingdom)

This document was typeset by the author using LaTeX 2$_\varepsilon$.

The jointly supervised research described in this book was carried out at the Laboratoire Informatique, Image et Interaction from the Universtité de La Rochelle and at the Centre de Visió per Computador from the Universitat Autònoma de Barcelona.

ISBN: XXX

Printed by Ediciones Gráficas Rey, S.L.

To my parents...

Nothing in life is to be feared,
it is only to be understood.
Maria Skłodowska-Curie (1867 - 1934)

Live as if you were to die tomorrow.
Learn as if you were to live forever
Mahatma Gandhi (1869 - 1948)

Acknowledgement

First of all I would like to thanks Dr. Simone Marinai and Dr. Apostolos Antonacopoulos to have reviewed this European thesis and for sharing interesting comments and discussions. I am also grateful to Pr. Dr. Koichi Kise and Pr. Dr. Jean-Philippe Domenger for accepting to be part of the jury and to Dr. Bart Lamiroy to have presided the jury.

Sincere thanks to my supervisors Pr. Dr. Jean-Christophe Burie, Dr. Dimosthenis Karatzas and Pr. Dr. Jean-Marc Ogier for their valuable expertise, friendship, dynamism, review of this thesis and joint supervision of this work between the Centre de Visió per Computador (CVC) of the Universitat Autònoma de Barcelona (UAB), and the Laboratoire Informatique, Image et Interaction (L3i) of the Université de La Rochelle (ULR). Also, thank you to Remy Mullot for accepting me in his lab (director of the L3i until 2012) and both institutions for providing me all the facilities to carry out this work in a great environment and atmosphere.

I would like to thank all the people from the Intelligent Media Processing Group of Osaka Prefecture University (Japan) for giving me the opportunity of doing research stays with a Manga's team: Pr. Dr. Koichi Kise, Dr. Motoi Iwata, Dr. Weihan Sun, the secretaries and all the master's students including Atsushi Ito and Masaki Tukada who have benefited of a research stay in France. I have learnt so many things about research, work, life style and food (thank you Hideto!).

Thank you to the CVC members for having welcomed me on several occasions (European doctorate and then joint supervision), especially people from the document analysis group: Josep Lladós, Dimosthenis Karatzas, Marçal Rusiñol, Oriol Ramos, Volkmar Frinken, Alicia Fornes, Jon Almazan, Lluis Gómez, Lluís Pere De las Heras, Anjan Dutta, Prass, David Fernández, Nuria Cirera, Francisco Cruz and Claire Pérez-Mangado from the secretary.

Thank you to the eBDtheque project's team for all the fruitful meetings we had together, Karell Bertet, Jean-Christophe Burie, Arnaud Revel, Alain Bouju, George Louis, Jean-Marc Ogier, Clément Guérin, Antoine Mercier (annotation tool and dataset's website), Norbert Tsopze and Nam Le Thanh. Thanks to the L3i lab members that participated to the ground truth day and also to the trainees that contributed to this project. Sincere thanks to the authors and publishers that have kindly agreed to share part of their works with scientists who allowed us to evaluate, share and make reproducible this three year thesis work. In alphabetical order: Pascal Boisgibault, Cyb, Fred, Sergio Garcia, Olivier Jolivet, Lamisseb, Gérald Lub-

bin, Winsor McCay, Midam, Marion Montaigne, Nicolas Roudier, Alain Saint Ogan, Trébla and Lewis Trondheim. Thanks also to their editors: Actes Sud, Ankama, Bac@BD, Clair de Lune, Dargaud, Delcourt, Doc En Stock, Dupuis, Hachette and Studio Cyborga. Finally, a special thanks to the CIBDI[1], Free Public Domain Golden Age Comics and the Department of Computer Science and Intelligent Systems of Osaka Prefecture University who kindly provided material from their personal collection.

Thank you to the people who have taught me how to teach, Vincent Courboulay, Anthony Bourmaud, Karel Bertet, Armelle Prigent, Arnaud Revel and Renaud Peteri.

I would like to thanks the European Doctorate founds of the University of La Rochelle, European Regional Development Fund, the region Poitou-Charentes (France), the General Council of Charente Maritime (France), the town of La Rochelle (France), the Spanish research projects TIN2011-24631, RYC-2009-05031 and the Sakura project of Campus France for the financial support of this work including the different research stays in Spain and Japan. Kind thanks to all the volunteer contributors of operating system, software, programming language and libraries related to the completion of this work (Fedora, NetBeans, Sublime Text, Gimp, Inkscape, Firefox, LaTeX, Python, IPython Notebook, Numpy, SciPy, Matplotlib, OpenCV).

I would like to thanks people that gave me the taste of academic research of image processing during the last year of engineering school and the research master final year project. You helped me a lot in the application process of this unique and wonderful thesis project, thank you so much.

Thank you to the members of the association of the Ph.D. students of La Rochelle (ADocs) for all the local events we organised together and the help for making and promoting a short film about my research work for young audience.

Grateful thanks to all the Ph.D. candidates, young doctors, engineers and administrative affiliated to the L3i with whom I spend most of the time (Romain, Clément, Guillaume, Cyril, Sébastien, Omar, Van, Giap, Sophea, Dounia, Maroua, Rouaa, Marcela, Bich, Imen, Hind, Phuong, Jane, Muzzamil, Benjamin, Joseph, Bruno, Olivier, Kathy, Caro, Sarah, Dom, Stéphane & Arnold and those I have forgotten sorry...).

Special thanks to the 121 bis office team (Nat, Mika, Toí, Elo, Nam, Bao, the Roundabout, the Cow and the Space invaders ^^) for these three years of gentleness, happiness and teasing that we shared together in this unique open space where science and creativity grow on the walls, roof and in the air.

And finally, I would like to infinitely thanks my parents, brother and all my family for their endless love and support at a distance of five hundred and twenty six kilometres as the crow flies.

Special thanks to my love for having followed me throughout this adventure with patience, dynamism, joy of living and also many contributions to this work (ground truthing, thesis review, etc.).

[1]Cité Internationale de la Bande Dessinée et de l'Image

Abstract

Born in the 19th century, comics is a visual medium used to express ideas via images, often combined with text or visual information. It is considered as a sequential art, spread worldwide initially using newspapers, books and magazines. Nowadays, the development of the new technologies and the World Wide Web is giving birth to a new form of paperless comics that takes advantage of the virtual world freedom. However, traditional comics still represent an important cultural heritage in many countries. They have not yet received the same level of attention as music, cinema or literature about their adaptation to the digital format. Using information technologies with classic comics would facilitate the exploration of digital libraries, accelerate their translation, allow augmented reading, speech playback for the visually impaired etc.

Heritage museums such as the CIBDI (French acronym for International City of Comic books and Images), the Kyoto International Manga Museum and the digitalcomicmuseum.com have already digitized several thousands of comic albums that some are now in the public domain. Despite the expanding market place of digital comics, few research has been carried out to take advantage of the added value provided by these new media. A particularity of documents is their dependence on the type of document that often requires specific processing. The challenge of document analysis systems is to propose generic solutions for specific problems. The design process of comics is so specific that their automated analysis may be seen as a niche research field within document analysis, at the intersection of complex background, semi-structured and mixed content documents.

Being at the intersection of several fields, combines their difficulties. In this thesis, we review, highlight and illustrate the challenges in order to give to the reader a good overview about the last research progress in this field and the current issues. We propose three different approaches for comic book image analysis relying on previous work and novelties. The first approach is called "sequential" because the image content is described in an intuitive way, from simple to complex elements using previously extracted elements to guide further processing. Simple elements such as panel text and balloon are extracted first, followed by the balloon tail and then the comic character position in the panel from the direction pointed by the tail. The second approach addresses independent information extraction to recover the main drawback of the first approach: error propagation. This second method is called "independent" because it is composed by several specific extractors for each elements of the image content. Those extractors can be used in parallel, without needing previous extraction. Extra processing such as balloon type classification and text recognition are also

covered. The third approach introduces a knowledge-driven system that combines low and high level processing to build a scalable system of comics image understanding. We built an expert system composed by an inference engine and two models, one for comics domain and another one for image processing, stored in an ontology. This expert system combines the benefits of the two first approaches and enables high level semantic description such as the reading order, the semantic of the balloons, the relations between the speech balloons and their speakers, and the interaction between the comic characters.

Apart from that, in this thesis we have provided the first public comics image dataset and ground truth to the community along with an overall experimental comparison of all the proposed methods and some of the state-of-the-art methods.

Résumé

Née au 19ème siècle, les bandes dessinées sont utilisées pour l'expression d'idées au travers de séquences d'images, souvent en combinaison avec du texte et des graphiques. La bande dessinée est considérée comme le neuvième art, l'art séquentiel, diffusé grâce aux progrès de l'imprimerie puis de l'Internet à travers le monde dans les journaux, les livres et les magazines. De nos jours, le développement grandissant des nouvelles technologies et du World Wide Web (la toile Internet) donne naissance à de nouvelles formes d'expressions s'acquittant du support papier pour profiter de toute la liberté du monde virtuel. Cependant, la bande dessinée traditionnelle continue a perdurer et représente un patrimoine culturel important dans de nombreux pays. À la différence de la musique, du cinéma ou encore de la littérature classique, elle n'a pas encore trouvée son homologue dans l'univers du numérique. L'utilisation des technologies de l'information et de la télécommunication pourrait faciliter l'exploration de bibliothèques en ligne, accélérer leur traduction et exportation, permettre de faire de la lecture augmentée (enrichissement du contenu lors de la lecture, à la demande et personnalisé) ou encore permettre l'écoute du texte et des bruitages pour les malvoyants ou les apprenants.

Les organismes de préservation du patrimoine culturel comme le CIBDI à Angoulême (Centre International de la Bande Dessinée et de l'Image), le musée international du manga à Kyoto (Kyoto International Manga Museum) ou encore le site digitalcomicmuseum.com aux États-Unis ont déjà numérisé des centaines d'albums dont certains sont du domaine public. Malgré la part de marché grandissante de la bande dessinée numérique dans les pays développés, peu de recherches ont été menées à ce jour pour valoriser ces contenus au travers des nouvelles technologies. L'analyse de document est une thématique de recherche qui facilite ce passage vers les nouvelles technologies. Une des particularités du contenu des documents est sa spécificité liée à son usage qui requiert bien souvent des traitements très spécifiques. Toute la difficulté de l'analyse automatique de documents réside dans la recherche de méthodes génériques capables de résoudre un maximum de problèmes spécifiques. Le processus de création d'une bande dessinée est propre à cet art qui peut être considéré comme une niche du domaine de l'analyse de document. En réalité, cette niche est à l'intersection de plusieurs problématiques de recherche qui compte les documents constitués d'un fond complexe, semi-structurés et avec un contenu varié.

L'intersection entre plusieurs thématiques de recherche combine leurs difficultés. Dans ce manuscrit de thèse, nous détaillons et illustrons les différents défis scientifiques liés à ces travaux de recherche de manière à donner au lecteur tous les

éléments concernant les dernières avancées scientifiques en la matière ainsi que les verrous scientifiques actuels. Nous proposons trois approches pour l'analyse d'image de bandes dessinées composé de différents traitements dont certains améliorent des travaux antérieurs et d'autres étant de nouvelles pistes d'exploration. La première approche est dite "séquentielle" car le contenu de l'image est décrit progressivement et de manière intuitive. Dans cette approche, l'extraction des éléments se succède, en commençant par les plus simples tels que les cases, le texte et les bulles qui servent ensuite à guider l'extraction d'éléments complexes tels que la queue des bulles et les personnages au sein des cases en fonction de la direction pointée par les queues. La seconde méthode propose des extractions indépendantes les unes des autres de manière à éviter la propagation d'erreur entre les traitements. Dans cette approche, les différents extracteurs peuvent être utilisés en parallèle puisque qu'ils n'ont pas d'interdépendance. D'autres éléments tel que la classification du type de bulle et la reconnaissance de texte y sont aussi abordés. La troisième approche introduit un système fondé sur une base de connaissance *à priori* du contenu des images de bandes dessinées qui permet d'interagir entre des traitements de bas et haut niveaux pour construire une description sémantique de l'image. Nous proposons un système expert composé d'un système d'inférence et de deux modèles sous forme d'ontologies, un pour modéliser le domaine de la bande dessinée, et l'autre pour modéliser les traitements d'images associés. Ce système, dirigé par les modèles, combine les avantages des deux approches précédentes et permet une description sémantique de haut niveau pouvant inclure des informations telles que l'ordre de lecture, la sémantique des bulles, les relations entre les bulles et leurs locuteurs ainsi que les interactions entre les personnages.

Dans cette thèse, nous introduisons également la première base d'images de bandes dessinées ainsi que la vérité terrain associée comportant des informations bibliographiques, spatiales et sémantiques. Cette base d'images annotées a été mise à disposition de la communauté scientifique. Des expérimentations basées sur les méthodes proposées et une comparaison avec des approches de la littérature sont également détaillées dans ce manuscrit.

Resumen

Nacido en el siglo XIX, el cómic se utiliza para la expresión de ideas a través de secuencias de imágenes, a menudo en combinación con texto y gráficos. El cómic está considerado como un noveno arte, arte secuencial, que ha incrementado su popularidad gracias a los avances en la impresión y el Internet. Hoy en día, el creciente desarrollo de las nuevas tecnologías y la World Wide Web da lugar a nuevas formas de expresión que permiten al papel disfrutar de la libertad del mundo virtual. Sin embargo, el cómic tradicional persiste y es un patrimonio cultural importante en muchos países. El uso de tecnologías de la información y de las telecomunicaciones podría facilitar la exploración de bibliotecas en línea, la traducción y acelerar su permiso de exportación, permitir una lectura con realidad aumentada (enriquecimiento de los contenidos durante la reproducción, a la carta y personalizado) o la escucha de texto y efectos de sonido para los alumnos o las personas con discapacidad visual.

Agencias de la preservación del patrimonio cultural como el CIBDI (Centro Internacional del Cómic y de Imagen) en Angoulême (Francia), el Museo Internacional de Manga en Kioto (Kyoto International Manga Museum) o el portal digitalcomicmuseum.com de los Estados Unidos han digitalizado cientos de álbumes, algunos de los cuales son públicos. Pese a la creciente cuota de mercado de los cómics digitales en los países desarrollados, poca investigación se ha llevado a cabo hasta la fecha para desarrollar estos contenidos a través de las nuevas tecnologías. El análisis de documentos es un campo de investigación que facilita a este problema de traslado a las nuevas tecnologías. Una particularidad de los documentos es la dependencia del tipo de documento que a menudo requiere un tratamiento específico. El reto de la análisis de documentos es de proponer soluciones genéricas para problemas específicos. El proceso de creación de un cómic es exclusivo de este arte que puede ser considerado como un nicho en el campo del análisis de documentos. En realidad, este nicho está en la intersección de varias tipologías de documentos ya que los cómics cuentan con un fondo complejo y un contenido semi-estructurado y variado.

Dicha intersección de varias tipologías de documentos combina sus dificultades. En esta tesis de doctorado, se describen e ilustran los diversos retos científicos de esta investigación con el fin de dar al lector toda la evidencia acerca de los últimos avances científicos en el campo, así como las barreras científicas actuales. Proponemos tres enfoques de análisis de imagen de cómic basados en diferentes tratamientos que mejora algunos trabajos previos y otros que son nuevas vías de exploración. El primer enfoque se denomina "secuencial" porque los contenidos de la imagen se describen gradualmente y de manera intuitiva. Objetos simples como cajas, texto y burbujas se

extraen primero y seguidamente se extraen las colas de las burbujas y los personajes de las cajas de acuerdo a la dirección apuntada por las colas. El segundo método ofrece extracciones independientes unas de las otras a fin de evitar la propagación del error entre aplicaciones, que es la principal desventaja del primer método. En este enfoque, los diversos extractores se pueden utilizar en paralelo, ya que no tienen interdependencia. Otros elementos como la clasificación del tipo de burbuja y el reconocimiento de texto están asociados. El tercer enfoque introduce un sistema basado en un conocimiento a priori del contenido de las imágenes de los cómics. Dicho sistema está basado en la interacción del procesamiento de imágenes a bajo nivel con la base de conocimiento para construir una descripción semántica de la imagen. Proponemos un sistema experto consistente en un sistema de inferencia y dos modelos ontológicos. Un modelo para el campo de los cómics y otro para el procesamiento de imágenes asociadas. Este sistema experto combina las ventajas de ambos enfoques anteriores y proporciona un alto nivel de descripción semántica que puede incluir información como el orden de lectura de los cuadros, las relaciones entre el texto y las burbujas, burbujas y sus personajes y la diferenciación entre los personajes.

Además, se presentan el primer conjunto de datos basado en imágenes de cómics públicos junto con su verdad terreno (que incluye la información espacial y semántica) que se ha puesto a disposición de la comunidad científica. También se detalla en este manuscrito un experimento de todos los métodos propuestos comparándolos con los métodos del estado del arte de la literatura.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter we provide the thesis context concerning its particular application to comic books. We remind the origin and societal role of comics in the different places in the world, the evolution from their creation to our digital word and its market place. Then the objectives and contributions of this thesis work are highlighted followed by the outlines of this manuscript.

1.1 Presentation

Comic books are a graphic art form combining text and images to tell a story. Comics are now used in a wide variety of styles, not only on paper (e.g. magazines, newspapers, TV show) but also as electronic content (mobile apps, e-books and websites). Comics are one of the most popular and familiar forms of graphic content. People read comics easily and learn many things, so even children can learn about cultures and trends, among other things, through comic books even unconsciously.

Comic art is difficult to define due to its intersection of several artistic mediums: graphic art, art film and the literature. More precisely it is drawing, film, writing, all combined together to form a new art (the ninth art) with an extremely varied set of expressions [46]. Wikipedia[1] defines comics as a visual medium used to express ideas via images, often combined with text or visual information. Comics frequently take the form of juxtaposed sequences of panels of images. Often textual devices such as speech balloons, captions, and sound effects ("onomatopoeia") indicate dialogue, narration or other information. Elements such as size and arrangement of panels and balloons control narrative pacing. Scott McCloud defines the comics as a "juxtaposed pictorial and other images in deliberate sequence, intended to convey information and/or to produce an aesthetic response in the viewer" [116].

There were early attempts to formalize the study of comics. Coulton Waugh

[1] http://en.wikipedia.org/wiki/Comics

attempted the first comprehensive history of American comics with *The Comics* in 1947 [198]. Will Eisner's *Comics and Sequential Art* in 1985 [49] and Scott McCloud's *Understanding comics: The Invisible Art* in 1993 [116].

The evolution of comic books The storytelling form that we know today as "comic book" goes back tens of thousands of years to the painting of animals, hunters, and shamans on caves walls for Christy Marx [27]. Nowadays, the history of comics must be described considering the three main centres of artistic creation. Europe where it is born, the United States to whom we owe its popularity and Asia that now represents the largest production of comics with Japan and Korea. Rodolphe Töpffer, a Francophone Swiss artist, is considered as the inventor of comics in the early part of the 19th century (1799-1846). He sequentially illustrated stories, with text compartmentalized below images and his art were reprinted throughout Europe and the United States. Magazine-style comic books emerged as a mass medium in the US in the thirties. From the fifties, Japan started a massive production of Japanese comic books called "manga" under the influence of Osamu Tezuka (comic artist) [21].

Laurence Grove mentions that current trends in French comics are at the harbinger of more general phenomena: globalization of critical traditions, acceptance of popular culture, blurring of borders between subject areas, a corresponding move away from a strict author-based canon, and high reliance on new technologies [60]. Magali Boudissa confronts the theoretical approaches developed around the paper medium to the changes brought by the story on the computer, focusing mainly on the management of a new space (screen and not a paper sheet) and the hypermedia aspect of the digital world [109].

Market place In France, after a period of crisis suffered by the sector during the years 1980-1990, the editorial production has enjoyed unprecedented growth, the number of publication was constantly increasing as well as the diversity of genres (e.g. manga, graphic novels, comics) [26]. In 2012, the French market reached the figure of 5,327 books published among which 72% were new (the rest being divided between re-editions, book art and testing). The annual sales reached 320 million Euro. In 2013, we saw a first drop of 7.3% that reflects the stabilization of the market and probably the emergence of the digital comics [150].

Japan is leading the sequential art market in terms of sales, about $5.5 billion in 2009 [179], and exportation volumes. In Europe and the Middle East the market is worth $250 million [41]. In 2008, in the U.S. and Canada, the manga market was valued at $175 million. Japanese culture is spread out all over the world through manga and anime. In 2014, the Japan Book Publishers Association published a handbook to take advantage of every opportunity to broaden and deepen overseas knowledge about the Japanese publishing world [33].

The US market is between Europe and Japan for single issues, collected editions or digital downloads, with $870 million for 2013 (up from $635 million for 2012)

according to the Comichron[2], a comics research site that tracks industry figures [125]. Digital sales rose from $70 million to $90 million.

1.2 Motivations

Nowadays, comic books, or "bandes dessinées" in French, represent an important part of the cultural heritage of one or two past centuries in many countries as mentioned above. Unfortunately, they have not yet received the same level of attention as music, cinema or literature regarding their adaptation to the digital format. Using information technologies with classic comics would facilitate the exploration of digital libraries [9], assist translators [16], provide a tool for augmented reading [76, 165], speech playback for the visually impaired [17, 145], story analysis etc. Nevertheless, the process of conversion and adaptation is not as simple as for films and novels. The comic differs from the latter in that the media itself is intimately linked to the medium. Indeed, a film can be decomposed into a series images plus a soundtrack. Just watching these images in the right order and at the right frame rate allows to reconstruct the initial content, regardless of the medium. In the same way a novel is ultimately a sequence of words. Reading these words in the correct order, on paper or on a screen does not change neither the content nor the artistic dimension of the work.

However, comics differ from the films on the form and the spatial positioning of the images. Where the latter pictures are all of equal size and each new image replaces the previous one, comic panels vary in size and spatial organisation in a limited space (paper sheet). These two features, added to the fact that the reader has the opportunity to see all the boxes of a same page, but not those of the next page, are tools at the service of the author to stage the story. Therefore, changing the medium, the reading surface format or the sequence order involves a modification of the staging that may in some cases be detrimental to the story. Moreover, space and time are closely related in comic art as demonstrated in Cortsen's thesis that explores the complexity of spatio-temporality and also focuses on how comics are structured in a network made up by individual elements and how they are connected [35].

The challenge of digital comics is how to take advantage of the added value provided by these new media such as personal computers, mobile devices and the Internet. Gilles Ratier, secretary of the French association of comics, critics and journalists "Association des Critiques et journalistes de Bande Dessinée" (ACBD) stressed in the 2013 budget of the association that the meaning of digital cartoon remains unclear [181]. It is commonly used to describe a wide range of heterogeneous content, from scanned comics to turbo-media works (e.g. interactive animated stories) through webcomics[3] (mainly self-published) and augmented reality to extend the art work through the use of the new technologies.

[2] `http://www.comichron.com/`

[3] `http://xkcd.com` or `http://phdcomics.com`

Web platforms propose free reading services of heritage art such as The Digital Comic Museum[4], the Cité internationale de la bande dessinée et de l'image[5] in France and the Grand Comics Database[6], an on-going international project that aims to build a detailed comic book database that will be easy to use and understand, and also easy for contributors to add meta-data to it in order to facilitate its browsing. In parallel, contemporary and independent artists self-publish their work on different platforms such as Webcomics.fr[7], EspritBD[8] and TRILLBENT[9] that recently proposed a promising per month subscription model. Well known publishers are also starting to provide digital comic book on pay-what-you-want hubs for original comic book platforms such as Koomic[10] and Panel Syndicate[11] from Spain, Izneo[12] and digiBiDi[13] from France, DC Comics[14] from Warner Bros, and comiXology[15] recently acquired by Amazon.com. In fact, we are seeing a bifurcation in the digital comics market, between companies tied to large global media conglomerates, that maintain a fervent faith in the need for some kind of digital rights management (DRM) control for their multi-billion dollar intellectual properties, and the smaller publishers more concerned with creator autonomy and exposure [102].

Others focus on innovative applications for mobile devices to provide a new reading experience by proposing their own reading application for mobile devices, integrating their technology. For instance Marvel[16] uses augmented reality and AveComics[17] from a French company Aquafadas proposes special transition and zooming effects that improve the reading experience through screens. Other initiatives, such as Sequencity[18], a project from the start-up company Actialuna in Paris, specifically focused on the ergonomics of the reading by transposing the bookstore experience into tablets (e.g. allowing rapid foliation and integrating communication systems with book sellers, able to provide personal recommendations). The connectivity to the Internet makes it also possible to enrich comic books with additional information from the web, allowing the reader to get extra content about the events, the places or comics characters related to the story being read. Note that in the rest of this manuscript "character" or "comic character" is used in the sense of actor or protagonist of the comics, not as a piece of text.

[4] http://digitalcomicmuseum.com/
[5] http://collections.citebd.org/
[6] http://www.comics.org/
[7] http://www.webcomics.fr/
[8] https://www.comixology.com/
[9] http://thrillbent.com/
[10] http://www.koomic.com/
[11] http://panelsyndicate.com/
[12] http://www.izneo.com/
[13] http://www.digibidi.com/
[14] http://www.dccomics.com/
[15] http://www.comixology.com/
[16] http://marvel.com/mobile/
[17] http://www.avecomics.com/
[18] https://www.sequencity.com/

These new uses generate technical needs related to several scientific issues when applied to large-scale processing. The International Digital Publishing Forum (IDPF) is currently working on a free and open e-book standard called EPUB 3 in order to make sequential art also benefit from the last advances of publishing technologies. EPUB 3 is a next-generation portable document format based on HTML5 and other Web Standards.

As mentioned above, several services propose printed to digital format conversion for comic books, mainly to facilitate reading on mobile devices for people that want to continue reading their favourite comics or manga on the way, without caring kilos of books. The need is increasing since the first generation of mobile devices with small screens in colour or B&W and now with smartphones and tablets. However, the conversion process remains tedious because done by hand (scanned and split into screen size parts small enough to avoid zooming and scrolling), and simplistic as it is too often reduced to the successive display of panels interspersed with user selected transitions. The ideal approach would be to understand the process used by authors to draw the paper-based comics and automatically transform it into a new form adapted to the medium in which the work is read (e.g. smartphone, web page, 3D book). This challenge is addressed by analysing the digitalized comic books in order to extract the different components (e.g. panel, balloon, text, comic characters) and their relations (e.g. read before, said by, thought by, addressed to). Once this initial work is done, it is necessary to reconstruct the story by retrieving the initial order of the extracted elements and also the links between elements in order to keep the story coherent. This is the role of the knowledge representation research field that addresses semantic understanding.

The computer analysis of comic book images is particularly challenging because they contain mixed contents of a graphical and textual nature. Furthermore, they are semi-structured documents meaning there is no regular structure allowing to extract easily the layout and to predict the location of text and graphics. Despite the fact that authors are entirely free in their layout choices, they follow few conventions widely adopted by comic book's authors in order to avoid the reader to be confused [46, 91]. The depicted elements and their place in the layout must be clearly identifiable at first sight, meaning, for instance, that balloons with speech text and characters should be included inside panels (location of the main actions). Comic documents are at the intersection between unstructured (e.g. teaching board [136], free-form document [43]) and complex background (e.g. advertising poster [29], real scene [50, 132, 199]) images which are nowadays active fields of research for the community. Being at the intersection of several fields of research increases the complexity of the problem. This is one of the reasons for which the analysis of comics is a recent (in the document analysis history) and not solved field of research.

1.3 Objectives and contributions

Comic books contain many heterogeneous elements that are difficult to process at once. Our objective is to retrieve as many elements as possible from any comic book image, using generic approaches able to handle such content heterogeneity. The final aim is to provide a high level description including spatial positions, inter-element relations and human-like interpretation of the images.

In this thesis we propose three approaches in order to cover the widest possible scope of study, from a sequential and supervised approach to an knowledge-driven and unsupervised method. The first approach profits from the relations between elements to guide the retrieval process. For instance, the panels are first extracted then balloons containing text that are inside panels and finally comic character regions of interest are defined from the speech balloon tail indications. This approach is quite intuitive but also very sensitive to error propagation issue between the different extraction steps [8,67]. The second approach consists in making the extraction independent from each other, in order to avoid error propagation issues. The third proposition adds contextual information to the independent extractions. The context is retrieved by matching extracted element relations with a generic model of the domain knowledge. This last approach allows a semantic description of the images. For instance, text regions that are detected inside balloons are inferred as being speech text regions and comic characters that are pointed by a tail as speaking characters.

We have constructed the first publicly available dataset and ground truth of comic book images to evaluate our contributions together with giving to the community the opportunity to work on reference data in order to make comparable and reproducible research. The last approach detailed above and the dataset are the results of a collaboration with Clément Guérin, a Ph.D. student working on knowledge representation and spatial reasoning applied to comic book contents.

To meet the above objectives, we have made the following contributions in this thesis:

1) Panel extraction: Comic books are mixed content documents that require different techniques to extract different elements. The first particularity of comics is the sequence of panels that we extract using connected component classification and topology analysis in Sections 3.2 and 4.2 respectively.

2) Balloon detection: Balloons or bubbles are key elements in comics; they link graphical and textual elements and are part of the comics style. They can have various shapes (e.g. oval, rectangular) and contours (e.g. smooth, wavy, spiky and absent). In this work we propose a method for closed balloon extraction and classification based on the analysis of the blob content (Section 4.4.1) and contour analysis (Section 4.4.2) respectively. A different approach for open balloons has been developed, it is based on active contour model to extract open balloons from text line positions (Section 3.3.2). Also, a tail detection and description are proposed (Section 3.4).

3) Text localization: Text can be of different nature in comics, there are sound effects (onomatopoeias), graphic texts (illustration), speech texts (dialogues) and narrative text (captions). Speech text represents the majority of the text present in comics, we propose an adaptive binary segmentation process followed by a text/graphic separation based on contrast ratio and then a text line grouping algorithm. Finally, an OCR system filters out non text regions (Section 4.3).

4) Comic character detection: Unsupervised comic character extraction is a difficult task as soon as we aim to process heterogeneous comic styles in order to cover all the comic books. In this context, learning-based approaches can not handle such dimensionality induced by the difference of styles. We first propose to define the region of interest of the comic character locations according to the contextual elements (e.g. panel contents, speech balloon position, tail position and direction) in Section 3.5. Second, we go one step further by spotting all the comic character instances in the album given an example and assuming that it has been digitized under the same conditions (Section 4.5).

5) Comics understanding: Enabling a computer to understand a comic strip is a really challenging task, especially because it is even hard for human sometimes (e.g. ambiguous reading order or speaker location). Putting comic domain knowledge in an ontology-based framework enables to interact between image processing and semantic information in order to progressively understand the content of a document (Chapter 5).

6) Dataset and ground truth: The eBDtheque dataset is the first publicly available[19] dataset and ground truth of comic book images. Such dataset is important for the community to make comparable, reproducible and growing research. The dataset consists in a mixture comic images coming from different albums with the goal of being as representative as possible of the comics diversity. The ground truth contains the spatial position of panels, balloons, text lines, comic characters and their associated semantic annotations. Also, bibliographic information is given for each image. This work is presented Section 6.1.

1.4 Outline

The rest of the thesis is organized as follows:

- Chapter 2 presents a detailed review of the state-of-the-art methods for the analysis of comic images. This chapter details several image processing methods in the four first subsections and then we review the holistic understanding systems that have been applied to document analysis so far. The last section reviews the more advanced services related to comics that are available on the market.

[19]Dataset website: `http://ebdtheque.univ-lr.fr`

- Chapter 3 introduces a sequential comic book image analysis approach. The sequence starts by extracting panel and text using connected-component labelling and clustering. Then balloons are segmented from text locations followed by tail position and direction detection on the balloon contour. The tail is used to define the region of interest for comic characters inside the panel regions.
- Chapter 4 addresses independent information extraction assuming no interaction between the extractions unlike Chapter 3. Different approaches are presented for panel, balloon, text and comic character extractions. Balloon type classification and text recognition are also discussed here.
- Chapter 5 presents a system that combines low and high level processing to build a scalable system of comics image understanding and enable a semantic description of the extracted elements.
- Chapter 6 presents the dataset and its associated ground truth, the metrics and an evaluation of the three proposed approaches. This dataset consists in a mixture images coming from different albums with the goal of being as representative as possible of the comics diversity.
- Chapter 7 concludes the thesis and defines the future directions of comic book document analysis.

Chapter 2

State-of-the-art

In this chapter we introduce the document image analysis principles and give an overview of the comics image analysis, holistic understanding of documents and the state-of-the-art of the applications related to comics in the society.

2.1 Document image analysis

A physical document is created when information for humans is transferred to a physical support (e.g. paper sheet, tracing paper, carbon paper, transparent sheet). This operation can be performed using various techniques and tools such as plume, pencil, pen, brush, stamp or machine-printed. Document image analysis is a sub-domain of computer vision systems that includes methods for acquiring, processing, analysing and understanding such physical documents in order to produce numerical or symbolic information (Figure 2.1).

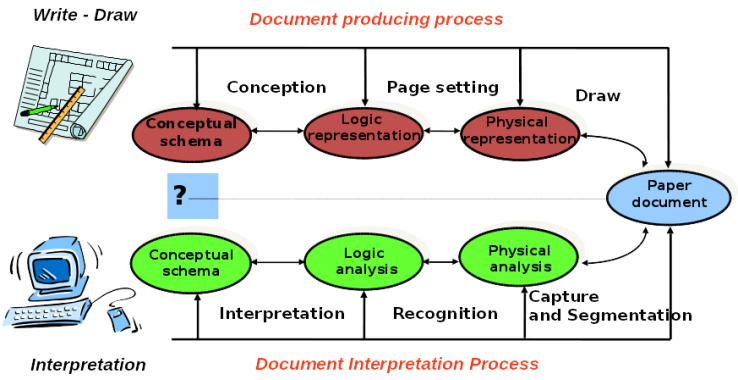


Figure 2.1: From the producer of the document to its interpretation [93].

The organization of a document analysis system is highly application-dependent in many cases, but many researchers worked and are still working on flexible and adaptable systems. Many functions are specific to the application, however, there are typical functions which are found in many computer vision systems.

- **Image acquisition** A digital image produced by the sensors of various devices (e.g. scanner, camera, web-cam) from the physical document. The digital image is an ordered set of pixels with a set of associated colour values from a given colour space.
- **Pre-processing** Before a document analysis method can be applied to digitalised image in order to extract some specific piece of information, it is usually necessary to process the image in order to assure that it satisfies certain assumptions implied by the set of processing methods used with the understanding process and its application domain (e.g. noise removal, contrast enhancement, segmentation) [12, 108, 138].
- **Feature extraction** Feature extraction is a process that consists in extracting relevant cues contributing to structure the information, especially for further processing. Image features are from different levels depending on their distance to pixel information (low level) and human understanding (high level). Low level feature such as line, corner, edges and blobs are computed directly from the spatial organisation of pixels [48]. However, high level information is usually computed from a global view of the image to extract texture and shape for example.
- **Detection/classification** At some point in the processing, decisions are made about which point of interests, symbols, textures, regions etc. are relevant for further processing. It aims to provide a first level of data organization and contributes to the recognition of elements of intermediate levels.
- **High-level processing** Integration of domain knowledge and final decision required for the application. This operation may require a user interaction that is sometimes taken into account for future automatic decisions.

The content of a document is generally divided into textual and graphical information. This corresponds to two active research fields, the first one, text recognition aims to convert the text into a character-encoding scheme such as ASCII. The second one, graphics or symbol recognition, focuses on the recovery of graphical information in documents. Graphics consist of spatial arrangements of symbols; examples include engineering drawings, maps, architectural drawings, music scores, formulas, tables, charts and some parts of the comic book image content (Figure 2.2).

Beside its content, a document is usually structured according to its intent of use. Tang mentioned in 1996 “A document has two structures: geometric (layout) structure and logical structure. Extraction of the geometric structure from a document refers to document analysis; mapping the geometric structure into logical structure deals with document understanding” [185]. Knowing the structure of the document

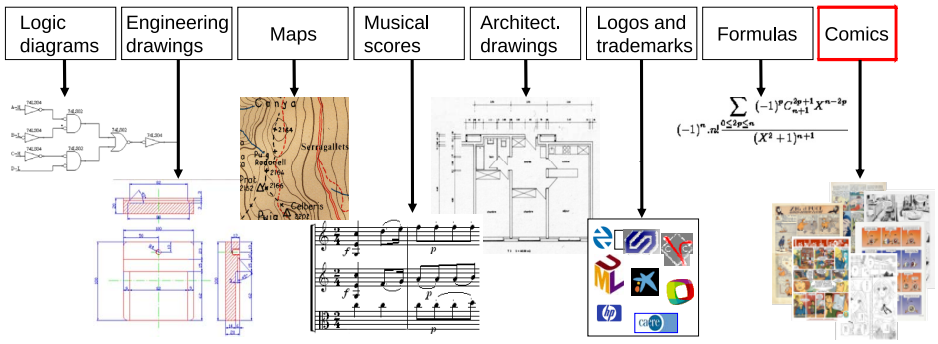


Figure 2.2: Application domains of graphics recognition [78].

being processed is always an helpful information that is related to domain knowledge information [18, 34].

Layout analysis consists in segmenting the image into several geometrical blocks that contain the same type of information: text, graphics, table, image, drawing etc. Then logical information can be retrieved using domain knowledge and the spatial position of the elements (e.g. header and title on the top, page number on the bottom-right corner, reading order). Layout analysis is not trivial for mixed content documents such as advertisements, posters and comic books. Those documents use non-standard text fonts, sizes and orientations mixed with graphics, images and logos.

Comic book images are composed of text and graphics that can be decomposed as drawings and line drawings. The diversity of texts that can be found in comic books images is large. From speech text which is mostly handwritten to sound effects that are sometimes close to drawing. Some of the graphics can be assimilated to symbols if we talk about the panels or balloons which have a sort of conventional representation among most of the comic albums. However, drawings contained in the panel regions do not follow any convention (free art) but contain repeated elements (e.g. comic characters) throughout the album or collection. This is related to comics' art that implies the drawing to be different from others in order to be easily distinguished and recognised by the public.

Comic images are mixed content documents with complex background, especially in the region of panels, that concerns the above mentioned field of research. Document image analysis being application-dependent, we briefly detail the design process of a comic image:

- **Synopsis and scenario** Imagine a story and its decomposition in a sequence of image (storyboard), view angles and format.
- **Pencil drawing** First rough drawing of the scenario, at this stage, the layout of the page if defined without any details.
- **Inking** The best pencil strokes are inked (permanent) for the final version.

- **Flatting and colouring** Here comes the colours (if any) into all stroke defined regions. Gradient, shadow and other effects are added according to the desired rendering.
- **Lettering and sound effects** Addition of text in reserved areas and sound effects over the graphics at the end.

The first challenge of comics book image analysis is to retrieve the layers that correspond to each step of the comic image design process (e.g. stroke, text, colour). Processing each layer separately would greatly simplify layout and content retrieval. Unfortunately, this decomposition is not obvious because in the final image, the elements are mixed with overlapping and transparency. Another important issue of document analysis is the noise information added from the creation of the physical document to its digital version. The main sources of noise are the manual operation variability, the degradation over time, the acquisition and storage techniques (Figure 2.3).

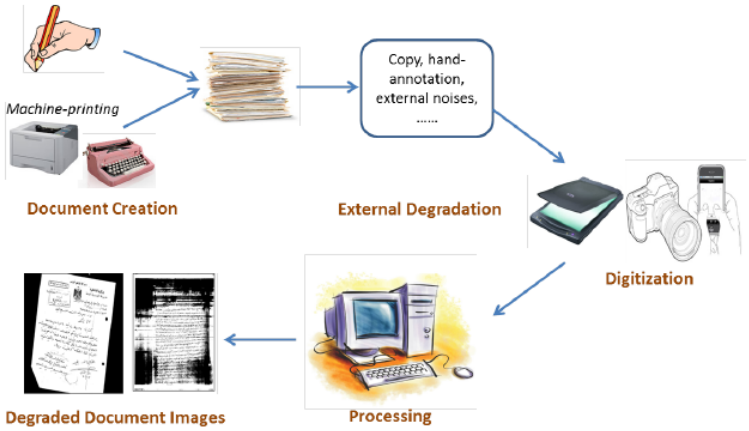


Figure 2.3: Document degradation sources [205].

2.2 Comic book image analysis

Comics images are mixed content documents combining textual and graphical information to create stories (Figure 2.4). Depending on the purpose, the document analysis techniques involved for comics image processing varies a lot between panel, balloon, text and comic character extractions. The contents of different natures are related to each other to produce a story (e.g. speech text is related to speech balloon and speaking characters). Treating each of the content separately has a limit that can be exceeded in a holistic understanding approach by reaching a higher level of semantic.

Figure 2.4: Examples of panels that reflect the content diversity of comic books.

2.2.1 Panel extraction and layout analysis

Panel extraction and ordering have been mainly studied for panel to panel reading. Several techniques have been proposed to automatically extract panels as [73], assuming that panels are small enough elements to be comfortably read on mobile devices. They are based on white line cutting algorithm [23, 47, 99], recursive X-Y cut [65] or gradient [184]. Those methods do not consider empty area [73] and border free panels. These issues have been corrected by connected-component approaches [7,139] but they are sensible to regions that sometimes connect several panels and increase the detection error rate. Another approach based on growing regions and morphological analysis can remove such connecting elements but also remove information on the panel border [67]. After the region segmentation step, heuristic filtering is often applied to classify the panel region according to the size ratio with the page size, which depends on the page format [8,67]. More recently, new methods have shown interesting results for manga and European comics with different background colours. They are based on watershed [146], line segmentation using Canny operator and polygon detection [99], region of interest detection such as corners and line segments [168,189]. Panel retargeting have been addressed for manga by Matsui [114].

Page layout analysis has been studied to calculate the reading order of the panels.

The page layout influences the reader at choosing pathway [32], nevertheless few studies [6, 61, 145] demonstrated the possibility of calculating such as Z-path (left-to-right and down) or right-to-left and down (e.g. Arabic, Japanese) according to the arrangement of the panels [100, 103, 189].

2.2.2 Balloon segmentation and tail detection

Balloons are a graphic convention used most commonly in comic books, comic strips and cartoons to allow words (and much less often, pictures) to be understood as representing the speech or thoughts of a given character in the comic[1]. Balloons are developed into a more-or-less oval shape, with a pointer or tail to indicate to which character they belong [55, 154]. There are many specialized forms of balloons, either traditional or invented [112]. See Figure 2.5.

(a) Oval/smooth (b) Rectangle/smooth (c) Oval/wavy (d) Oval/spiky (e) Suggested

Figure 2.5: Example of speech balloon types by *shape/contour* types. Image credits [38, 74, 115].

In comics' content understanding, speech balloons present a lot of interest since they offer the links between the textual content and the comic characters providing information about the localization of the characters and the tone of speech. Apart from being crucial for document understanding, balloon detection is also beneficial in applications such as comic character detection [174], content re-targeting [114], translation assistance and reading order inference [61].

Few works about balloon extraction have been done until now and mainly closed speech balloons have been studied. Arai [8] proposed a blob detection method based on connected-component detection with four filtering rules applied to manga analysis. The rules are based on blob minimum size, white pixel occurrence, inclusion of vertical straight lines and width to length ratio (Figure 2.6). Another connected-component approach proposed by Ho [67] uses HSV colour space to make a first selection of bright blobs and then consider as balloons the blobs with a ratio between the text area and the blob bounding box higher than sixty per cent.

The analysis of partially closed balloons (also called open, implicit or suggested) is quite different than the closed ones which are mainly based on blob analysis (Figure 2.5e). From our knowledge, such type of balloons has not been studied before. In

[1] `http://en.wikipedia.org/wiki/Speech_balloon`

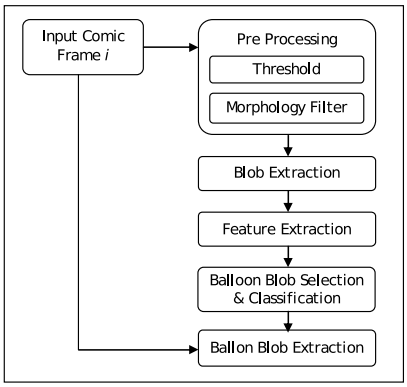


Figure 2.6: Flow diagram of comic balloon detection using comic blob extraction method proposed by Arai [8].

most of the cases, including when contours are implicit, the location of text is generally a good clue to guess where the balloon is. The problem of speech balloon outline detection can therefore be posed as the fitting of a partially drawn contour around text areas (Section 3.3.2). Deformable models such as the active contour [83] model, also known as snake, try to minimize an energy term associated with the initial contour and its neighbouring elements in order to shift it to the position of the suggested contour. Kass [83] proposed the original energy functions able to shrink around an object contour, Xu [200] proposed the Gradient Vector Flow external force to attract snake from further and handle broken object edges and subjective contours. Another extension was proposed by Cohen [30] to make the curve behave like a balloon which is inflated by an additional force. Finally, active contour in a multi-resolution context have been studied to speed up the process on multi-resolution image and the multi-resolution model itself [96].

Balloons provide also extra information about speech tone according to the different patterns which are along the contour of the balloon. The shape of the balloon does not provide a lot of information about how the text is spoken; it is more related to the style of the comics and the structure of the panel. Therefore, we focus on contour classification methods. In the literature, contour classification is strongly related to shape classification purposes [104, 171]. It has been applied for video [10, 89, 152], trademark retrieval [97], speech recognition [59]. Also, wavelet decomposition [186] and invariant moments [130].

From our knowledge, tail detection has not been studied before, we propose to analyse the contour patterns in order to locate the tail and then perform a local analysis to find its direction (Section 3.4).

As far as we know, balloon classification has not been studied before. It is related to shape and contour classification field of researches in planar images. Shape classification is a well developed field, especially in a template matching context. Three well know methods are Curvature Scale Space (CSS) [1], Fourier Descriptor [207] and In-

variant Moments [71]. A recent work [105] explains that CSS is more appropriate for shapes containing a high number of inflections as marine animals [1]. Apart for document analysis, shape classification was also studied for image retrieval such as marine creature [128], leaves [197], illicit tablet [105], anthropology [85] and visual shape descriptor [15]. A good shape representation overviews are given by [192] and [208]. Nevertheless, shape descriptor and classification are not discriminant enough for our purpose because speech balloon are most of the time compact, symmetric and hand drawn which corresponds to limitations of the reviewed methods.

Note that balloon classification and tail detection are also important in a context of dialogues and emotions understanding [126].

2.2.3 Text extraction and recognition

The challenges of text extraction and recognition in comic books are the multi-script, multi-oriented and complex background aspects. Figure 2.7 illustrates the diversity of text in comic books. First of all, the text lines are quite short compared to other types of documents (see experiment Section 6.1.1) and, according to the style of the comics, there are variations of fonts (mainly handwritten), case, orientation, scale, spelling (from dictionary or with voluntary spelling mistakes) and hyphenation.

Figure 2.7: Examples of text lines from comics images.

Document analysis techniques are not designed to handle such diversity of texts within the same document; we first review another related field of research which is scene text detection and recognition. Text localization in real scenes and video frames is an active research topic [79, 81, 161]. Recent methods have been propose for scene text detection [56, 82]. However, applying existing real scene text detection methods to comics would not be optimal to cope with all the different types of text that are combined in comic book documents (e.g. typewritten, handwritten, graphic

sound). If we consider typewritten text, the most similar application to comics is car plate recognition because the text is in a salient and contrasted area with a complex background around the plate such as speech balloon for comics [5].

The documents that have attracted of lot a attention are newspapers, administrative documents, cheques, maps, music scores, floor plans and engineering drawings [131]. Also, an important effort has been done to separate text from graphical content from such types of document with the focus on improving the post processing of both types of content by treating them separately. Fletcher and Katsuri [54] still the reference in this domain more than twenty years after for more and more complex documents [155, 188]. From our knowledge, only one preliminary study has been proposed for text/graphic separation in comics [11]. This approach relies on artificial neural networks, edge and corner detection. The authors report results on very few panels and only for speech text, thus it can not be generalised to that we can find in comic book images (e.g. speech texts, onomatopoeias, captions). Nevertheless, text localization in comic images have been recently investigated and opens up several interesting applications such as image compression [169], OCR training and also translation, speech synthesis and re-targeting to different mobile reading devices [114].

Comics being semi-structured documents mixing textual and graphical information, they combine the difficulties of both domains, making the task of text localization especially challenging. Text localization in complex images has been previously studied in scene images [50, 121, 132, 195, 199], video sequences [87, 163] and digital-born images (Web and email) [80]. Text localization in unstructured documents has been studied for teaching boards and slide show presentation [133, 136, 191]. However, text localization in documents which are both unstructured and have complex background has received relatively little attention [29].

Few works concern text extraction in comics, they all rely on speech balloon regions. Bottom-up approaches use connected-components which often relies on the segmentation step [146]. Su [169] uses Sliding Concentric Windows as text/graphic separation and then apply mathematical morphology and a Support Vector Machine (SVM) classifier to classify text from non-text components. Li [101] proposed unsupervised speech text localization for comics that trains a Bayesian classifier on aligned connected-component and then detect the rest of the text using the classifier for text/non text separation.

Top-down approaches starting from balloons (white blobs) detection followed by mathematical morphology operations have been proposed by Arai [8], Yamada [203] and Sundaresan [177]. From our knowledge, there are no published work concerning graphic sounds (onomatopoeia) and illustrative text extraction.

Text recognition applied to comics is really challenging because it includes most of the difficulties from text recognition in document analysis domain if we consider all the types of text that compose the comics. From typewritten to handwritten, free-form text in uniform to complex background including image noise, text deformation and overlapping. Nevertheless, Ponsard [146] solved a sub part of the problem by focusing on speech text of a single typewritten font and language for which an OCR

system is trained for.

2.2.4 Comic character detection

Human detection in computer vision field has been largely studied during the past decades, mainly based on greyscale image and gradients. Colour information is rarely relevant for human detection because of clothing and skin colour difference. In videos, moving regions are often used as region of interest.

Although, comics are often a reproduction of human life situations, it is a domain where we can not directly apply human detection based methods. The main difference is that comic characters (e.g. protagonist, hero) are hand drawn and therefore more variant in terms of deformation, shape and appearance than real life humans (Figure 2.8). In coloured comics, the colour information gives the identity of characters and plays a main role for character spotting with speech balloon positions. The simplistic character design allows for easy identification and representation and goes along with how human process visual information [3, 31, 119]. This is a big difference compared to natural images and human detection, since comics are designed in a way that the information (e.g. where are the comic characters, who is talking, where is it going) can be quickly found.

Figure 2.8: Examples of comics character postures. This example shows deformation, pose, rotation, translation and occlusion variations. Image credits: [147].

Recent studies have been published for partial manga copy detection [173], mainly based on shape information because of the absence of colour information. This work has been extended to Manga copyright infringement protection [172, 175] using Maximally Stable Extremal Regions (MSER) [113] or faces [193] for region detection and Histogram of Oriented Gradients (HOG) [40] as region descriptor. A recent study discusses the local feature extraction and the approximate nearest neighbour (ANN) search [75] . This work shows good results for manga part retrieval which is a subset of the whole comics world.

Coloured comics may be compared to cartoon images sequence, for which a first work based on HOG, SVM and colour attributes has been published in 2012 [86]. Preliminary work about cartoon and comics faces recognition has been carried out by Kohei [183], Cheung [24] and Bernhaupt [11]. More recently, graph theory has been used to find redundant colour structure in order to automatically localize the most frequent colour group apparitions and label them as main characters [68]. The thesis of

Ta [182] (sections 1.1 and 1.2) gives a good overview about stroke-based image analysis similar to comics, it defines the issues of poor information, occlusion, deformation, inter-class and intra-class variations, scale, spatial and/or temporal relations and structured data. Ta [182] also mentions that "story board scene understanding still remains an open problem, few results are available in the literature about stroke images".

Another work uses HOG descriptor with redundant information classification to also find the most frequent elements [176]. Both graph and descriptor based methods need an image pre-processing step to remove irrelevant redundant elements such as text and speech balloons.

Other interesting approaches try to automatize comic generation [187, 196] using image cartoonification and script analysis.

2.3 Holistic understanding

One of the original goals of image or graphical document analysis was to fully understand the content of any image [93]. This requires solving several sub-tasks simultaneously, for instance region detection, labelling of meaningful regions and semantic understanding using layout analysis. In the past, researchers have developed classifiers for tackling each of these sub-tasks independently [111]. However, these sub-tasks can help each other but it is rarely easy to combine different related sub-tasks. For instance, in a comics page, if we know the panel positions, then we can make a better guess at the location of the comics characters (because they are usually inside the panels). Previous works concern real scene image analysis [14], retrieval [159] and understanding [53, 98], medical image annotation using description logic and inference engine [70], object-based image retrieval [123, 158] (between keyword-based and query-by-example) and image interpretation [72, 135] that mentions the importance of using topological information, distances, directional relative position and much complex relations such as "between" and "surround" and "among". Stroke-based images such as engineering drawings, electronic circuit diagrams, architectural drawings and floor-plans have also been investigated [45, 140, 141]. Also, Geographic Object-Based Image Analysis (GEOBIA) [14] makes extensive use of ontologies to interpret maps.

Recently, comic book images have been also considered. An ontology of comics has been proposed from a philosophical approach [122], a semantic annotation tool [66] makes use of previous knowledge and consistency information to suggest new knowledge to the user in a interactive way. Spatial inferences have been used to infer the comic books reading order, for panel in the page and balloon in the panel [61]. In [159], they highlight the benefit of using contextual information of simple objects to build more complex ones.

2.4 Existing applications

We can divide comic-related computer systems into two categories. One type would be using comics to represent complex information such as online communication in a form of comics [90] and generating video or story log summaries [4, 162, 190] (computer graphics domain). Also, it is possible to listen to manga [164] which have been recorded by people reading the story or to use mobile app for automatic translation [19, 20] (requires user text selection). These systems are useful to add value to the content while making it more funny and interesting. The other category concerns the comic design by enabling novices to create comics interactively using a computer for augmenting an individual user's memory [170], turning photo albums into a comic [28, 143], making comic-like video [151], making collaborative comics [106] and exchanging rich messages [157].

2.5 Conclusion

Some of the challenges of comics image analysis can be highlighted from the above state-of-the-art reviews. First, comics images suffer from noise as any other document image processing. It comes from hardware processes (e.g. drawing techniques, printing and digitization) and software (e.g. image compression). So efficiently handling noise is crucial for image analysis and understanding. Knowing the design process of the comics creation helps for image denoising. Second, according to the results of the reviewed methods, we can order them by level of difficulty, from the simplest to the hardest: panel, balloon, text, comic character and holistic understanding of an image or album. Third, most of the works in the literature use different copyrighted images which can not be shared publicly. Moreover, the authors usually do not share their code. This is a key issue for researchers which can not share, reproduce and compare results on identical data in a collaborative way. This is one of the reasons that retains comics analysis to progress as fast as other field of research of document analysis.

In the next chapter, we are going to present a sequential content extraction approach that profits from the relations between elements to guide the retrieval process. We first apply panel and text region extraction followed by speech balloon and tail segmentation. Then, speech balloon tail indications are used to compute a comic character region of interest according to the spatial organisation of previously extracted elements in each panel.

Chapter 3

Sequential information extraction

This chapter presents a sequential information extraction approach for comic book content retrieval. The sequence of extraction starts from elementary elements such as panel and text that initiate further processing. Once the text regions are discovered, we use them as seeds to search for balloons and then we analyse the balloon contours to detect the tails. From the tail position and location, comic character regions are computed according to previously extracted element positions.

3.1 Introduction

In comics art, the page structure depends on the author style, this is why so many different structures and drawing types exist. This author style (a kind of signature) gives a unique graphical identity to the comics and contributes to attract the curiosity of the readers. Despite the differences of style, the comics drawings generally follow classical rules which are intrinsically linked to the design process [117]. For instance, the inking process requires to magnify the main strokes that defines regions which are then filled with flat colours (Section 2.1). We propose to rely on these main strokes to automatically extract non-coloured image content (e.g. panel boundary, text, speech balloon background) using a connected-component labelling approach. Then, the analysis of speech balloon contour allows detecting the tail precisely and its direction is used to compute the regions of interest where comic characters are supposed to be. In this section, all the processes are related to each other, for instance, the text extraction output is the input of the balloon extraction process and so on.

This is a simplistic and intuitive approach to identify the benefits, difficulties and limits that result from this type of approach. It will be associated with an intelligent system in the knowledge-driven approach presented in Chapter 5.

3.2 Panel and text

This section proposes a method to automatically extract the panels and text regions contained in comics pages. This method is not limited to text which is included into speech balloon such as most of the work in the literature. Here we consider all the text regions of the image using connected-component labelling approach and k-means clustering.

To be clustered, the connected-components have to be extracted from the image first. A commonly used method is to segment the original image pixels into two categories called foreground and background. The foreground category corresponds to the set of pixels of interest, here the pixels of the panel border strokes and text letters. The background includes other pixels. This is usually performed using binary segmentation techniques which assign each pixel to one of the two categories. From the foreground pixels, a structural analysis (connected-component labelling) allows us to group "connected" pixels into components according to their connectivity.

Connected component approaches are commonly used for stroke and text-based document analysis, they are also simple and computationally efficient. Once extracted, the connected-components can be clustered according to different features (e.g. size, shape, colour and location). We use this approach to extract panels and text which can be easily differentiated using size and topological information. The process can be summarized as follows:

1. Binary segmentation
2. Connected-component extraction
3. Connected-component clustering
4. Candidate regions pruning

Binary segmentation

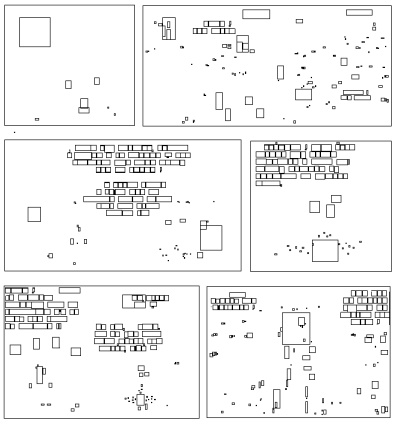
The first step consists in a colour to grey level (3 to 1 channel) conversion. Several methods exit for such conversion, combining the three channels of the red, green, blue (RGB) colour space with different pre-weighting as given in [148] or using one channel from the Hue Saturation Lightness (HSL) or Value (HSV) representation of the RGB colour space. Here we use the lightness channel because the panel border strokes and text are usually darker than other elements in the image, including the background. Then, a global binary segmentation (Appendix A.1.4) is applied with a threshold determined dynamically (Figure 3.1a). The threshold is computed from the median value of the border page pixels where pixels with a value lower than the median value are considered as part of the foreground (we are interested in black strokes). We assume that the border pixels of the page are representative of the page background (depending on the digitization process). If the median value is closer to "black" than "white" grey level, then, image inversion is applied and we redo the complete process in order to always get a white background at the end of this step.

The binary conversion step is very important for the rest of the method because the background part will not be considered for further processing.

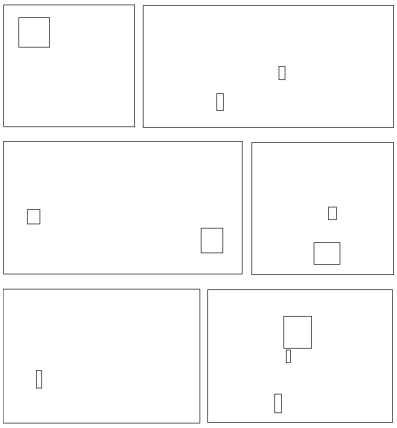
Connected-component extraction

The connected-component algorithm (Appendix B.1) is applied on the binary image and the bounding box of each black component is computed to facilitate subsequent processing (Figure 3.1b). We do not consider white connected-component here because we assume that panel border and text elements are darker than other elements in the page (segmented as black region).

(a) Binary segmented image



(b) Connected-component bounding boxes



(c) Panel cluster

(d) Text cluster

Figure 3.1: Panel extraction process. Image credits: [38].

Clustering

By looking at the figure 3.1b, we can clearly see that the bounding boxes of the panel are bigger than the others. Also, they include more regions than others. Classifying those regions according to their size allows us to classify panel and text region at the same time while ignoring noise information. Knowing the number of clusters facilitates the clustering, here we set the number of clusters $k = 3$ according to the domain knowledge of comics. They aim to reflect the "panel" (the highest), the "text" (medium height) and the "noise" (few pixels height) as shown on Figure 3.2.

To perform the clustering, we use one of the most popular clustering algorithm which is k-means clustering method (Appendix A.1.5). The clustering is performed dynamically on each image which makes this method parameter-free (for a given number of clusters) and invariant to page format and resolution.

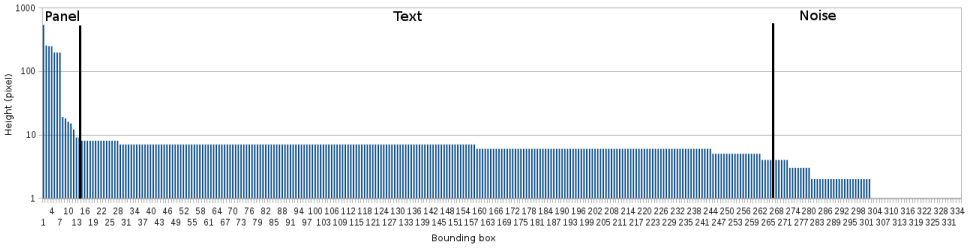
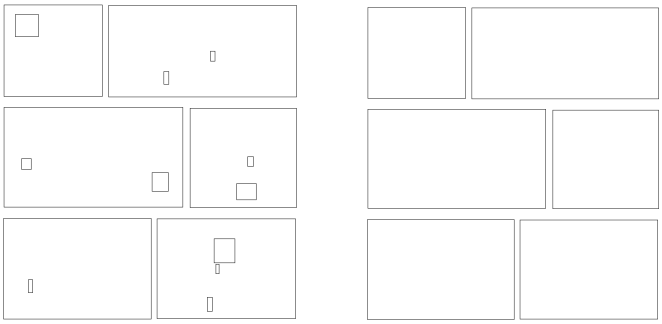


Figure 3.2: Descendent histogram of the bounding boxes in figure 3.1b. Vertical black lines represent an example of three cluster limits labelled as "panel", "text" and "noise". Note that the lowest part of the "panel" cluster do not correspond to panel regions, we present how we pruned them in the next paragraph.

This method assumes that the image is highly contrasted and contains a high disparity between stroke sizes otherwise the binary segmentation or the clustering process may fail.

Panel pruning

The results from the clustering operation can be pruned using domain knowledge. The domain knowledge is given by the most followed conventions of comics design such as the panels that are usually juxtaposed and rarely included into each other 1.2. Note that the integration of the knowledge directly into the low level pruning process is restrictive for future processing that will have no chance to recover missing information. Based on this assumption, we take out components which are included in other regions of the panel cluster. Given the set of component bounding boxes from the panel cluster $R = \{R_1, R_2, ..., R_n\}$, we filter out panel candidates that do not verify this relation $R_i \notin R_j \forall j, i \neq j$ (Figure 3.3a and 3.3b).



(a) Candidate panels (b) Pruned panels

Figure 3.3: Topological pruning of panel bounding box extraction.

Letter to text line

The gap between text component (e.g. letter or attached letters, word) and text lines can vary significantly. In fact, handwriting generates a lot of artefacts such as alignment variations, mergers between letters and text lines. We propose a simple method that handles the two first aforementioned artefacts considering the height and the neighbourhood of the connected-components.

Among all the connected-components (Figure 3.1b), we clearly see that the spatial organisation and the regular size are characteristics of text regions (regardless of the language). We group text candidates into text lines according to alignment and ignore text candidates that can not be part of any text line.

Similarly to [29, 100], we first search for the first letter (connected-component) of each text line and then attach the close and similar connected-components to the same text line. A letter is considered first if it is positioned on the left, on the same horizontal line and if there is no intersection found with any other letters at a distance equal to the letter height. Then the other letters on the right are added by checking their relative horizontal and vertical positions. For this purpose, we defined two conditions that are auto-adaptive to each letter (Figure 3.4):

- The horizontal inter-letter distance d should be smaller than the maximal height of the letters ($d < max(h_1, h_2)$);
- The vertical alignment is considered as correct if the vertical coordinate of the centre of the next letter c_2 passes through the first letter ($y_{min}(letter_1) > c_{2y}$ and $y_{max}(letter_1) > c_{2y}$);

The principle is similar to the one presented in [29] but adapted to take into account the horizontal alignment of text lines. Our method does not consider the letter width as we never know how many letters correspond to the CC. This method can easily be used for vertical text (e.g. Japanese, Chinese and Korean) by switching

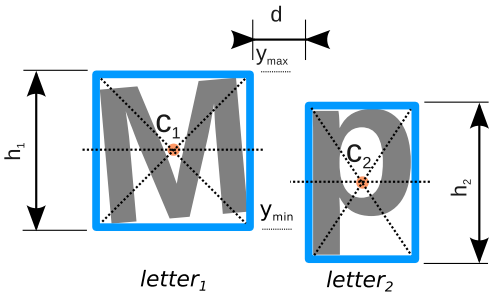


Figure 3.4: Text component horizontal and vertical alignments. The two rectangles represent the bounding boxes and c_1 and c_2 their centres.

horizontal and vertical measurements. More text orientation can be handled using automatic multi-oriented touching character segmentation such as [156].

Text line to paragraph

We post-process the result of the text line detection to group text lines into text area (paragraph), according to two rules. First, we require that the candidate text lines to group have similar heights (or width in case of vertical text) and second that the inter-line distance is smaller than the average text line height of the potential paragraph region (Figure 3.5).

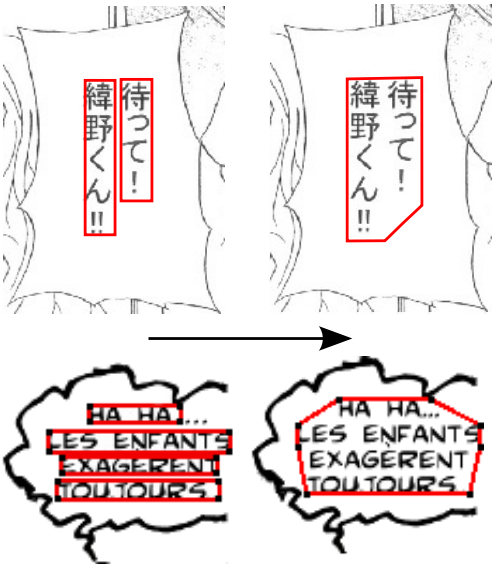


Figure 3.5: Illustration of the text line to text paragraph conversion from left to right for vertical and horizontal text. Image credits: [38, 74]

3.3 From text to balloon

For comics content understanding, speech balloons are of significant interest since they offer the links between the textual content and the comic characters providing, in this way, information about the localization of the characters and the tone of speech. As mentioned in Section 2.2.2, speech balloons are the more frequent type of balloon in comics and are highly related to speech text regions. In this section, we propose two approaches for balloon extraction based on text region location detection using the methods presented in Section 3.2 and 4.3. The first method defines as "balloon" the smallest connected-components that contain several text-like regions. The second method groups text lines into paragraphs and initializes an active contour model (snake) on its outline. From the initial position, the snake is pushed away from the text and attracted by surrounding edges at the same time in order to stick to any potential balloon stroke close by. The main difference with the first approach is its ability to detect implicit or partially drawn balloons as well as the closed ones.

3.3.1 Regular balloon extraction

Regular balloons are defined as closed balloons with a completely drawn contour unlike implicit contour discussed in the next Section 3.3.2. Closed balloons are easily extractable using blobs detection method similarly to the panel and text extraction presented above. The main difference lies in their dominant colour which is generally white and implies to extract white connected-components instead of black ones as for panels and text. One particularity of text inside balloons is its vertical and/or horizontal alignment in the balloon (Figure 3.6). We propose to use this characteristic to compute a confidence value to each connected-component that includes one or more text regions.

Figure 3.6: Example of text positions in speech balloons. Image credits [38, 74, 115].

Blob extraction

A typical "closed" balloon is surrounded by a black stroke that defined a white region (blob) inside it, with holes created by the presence of text letters. We propose to

make a binary segmentation of the comics image and then use connected-component labelling method to extract white blobs (Figure 3.7). We assume that closed balloons are surrounded by a stroke with similar darkness and stroke-width from one balloon to another in a given scanned image. In this context, global segmentation methods are appropriate because the optimal threshold selection is similar for all balloon regions. We use Otsu's threshold selection method to find the optimal threshold (Appendix A.1.4).

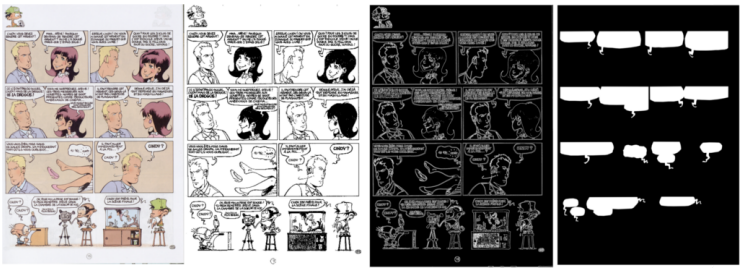


Figure 3.7: Balloon extraction process. Original image, binary segmentation, contour detection (connected components) and result mask from left to right. Image credits: [124].

Balloon extraction

For each extracted blob that includes a text paragraph (Section 3.2), we compute the difference of alignment on the horizontal d_x and the vertical d_y axis between the balloon barycentre c_1 and the text paragraph barycentre c_2 (Figure 3.8). We use the difference of alignment as a confidence value $C_{balloon}$ for balloon and non-balloon classification (Chapter 6).

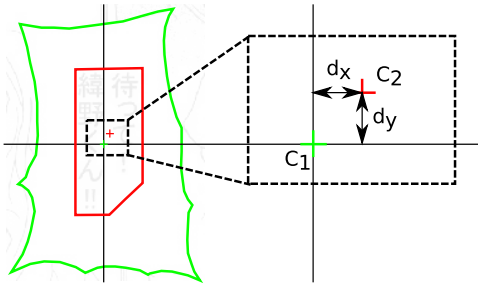


Figure 3.8: Vertical d_y and horizontal d_x alignment differences between balloon barycentre c_1 and text region barycentre c_2.

Both differences d_x and d_y are normalized between zero and one as a percentage of the balloon width B_{width} and height B_{height} respectively. The average of the vertical and horizontal alignments gives a confidence value for each balloon candidate (Formula 3.1).

$$C_{balloon} = 1 - \frac{1}{2} * \left(\frac{d_x}{B_{width}} + \frac{d_y}{B_{height}} \right) \tag{3.1}$$

If the sum of differences of alignment is higher than the balloon size ($d_x > B_{width}$ and $d_y > B_{height}$) then $C_{balloon}$ becomes negative which is considered as $C_{balloon} = 0$. An alternative could be to compare the axis of the main moment of inertia of text and balloon regions.

3.3.2 Implicit balloon extraction

Balloon contour is not always completely drawn, it may be implied by contextual information such as contrast difference or other surrounding elements (Figure 3.9c). In most of the cases, the location of text is a good clue to guess where the speech balloons are. The problem of speech balloon outline detection can therefore be posed as the fitting of a closed contour around text areas, with the distinctiveness that the outline might not be explicitly defined in the image. For the examples given in Figure 3.9, this would be a smooth contour with relatively constant curvature (Figure 3.9a), an irregular one with high local curvature (Figure 3.9b), and an implicit one with missing parts (Figure 3.9c). Through these observations, we can see how important the domain knowledge is, in the global interpretation process.

a) Smooth b) Irregular c) Implicit

Figure 3.9: Example of speech balloons with different contour types. Image credits [38].

Observing the heterogeneity of balloons, and considering the difficulty to manage "open" or "implicit" balloons, it appears necessary to use a dynamic and adaptive outline detection algorithm. Active contours appear to be suitable to the problem. The active contour framework was developed for delineating an object outline in an image. The algorithm attempts to minimize the energy associated to the current contour, defined as the sum of an internal and an external energy term (Appendix A.1.3). In this section we propose two adaptations of the active contour theory to the domain of comic balloon detection. Specifically, we handle the case of balloons with missing parts or implicit contours, while we adopt a two-step approach to fit irregular outline types such as peak and cloud type balloons. To achieve this, we propose new energy terms making use of domain knowledge.

The introduction of statistical shape knowledge has already been studied in the literature but can not be applied here because of the lack of knowledge about the contour shape to detect and complex background [36]. The original active contour model proposed by Kass [83] defines an energy function composed by internal E_{int} and external E_{ext} forces that make the initial contour shrink to the object boundaries (Appendix A.1.3). As we assume that text is included inside balloons, we initialize the contour model on the text block boundary position. In our case, we need to inflate instead of shrink the initial contour in order to retrieve or approximate the balloon boundary. For this purpose, we introduce a new energy term denoted E_{text} that conveys information about the relative placement of the balloon outline and the enclosed text (Equation 3.2). Note that for the discussion below, we assume that text has been already detected in the image (Section 3.2 and 4.3).

$$E = E_{int} + E_{ext} + E_{text} \tag{3.2}$$

External energy

We consider edges as features of interest because we expect the speech balloon to be delimited by strong edges, at least partially. Here we perform edge detection using the well-known Sobel operator (Figure 3.10). According to the review of state of the art methods appropriate for implicit balloon detection (Section 2.2.2), active contour models are good choices. Appendix A.1.3 gives the basic concepts of the original approach proposed by Kass [83]. In this approach, the definition of the external energy is appropriate for natural scene images with smooth gradients but not for stroke-based images such as comics that comprise uniform coloured regions (flats) with strong edges (strokes). In our case, we require that edges attract the snake from relatively far away (distances where the original edge gradient has already dropped to zero). The method of Xu [200] would be appropriate here, although we have decided to use the equivalent distance transform of the edge image instead for computational efficiency reasons. We therefore define the external energy function as:

$$E_{ext} = \gamma \min A(i,j) = \gamma \min \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{3.3}$$

where E_{ext} is the minimum Euclidean distance (A) between a point i and the nearest edge point j, γ is a weighting parameter.

Since it is not desirable for edges corresponding to text to attract the snake, any edges that fall within the text regions are removed before computing the distance transform and do not contribute to the external energy.

Internal energy

We use the original definition of the internal energy (Equation A.1) which can by decomposed in two energy terms: $E_{cont} = |\mathbf{v}'(s)|^2$ and $E_{curv} = |\mathbf{v}''(s)|^2$.

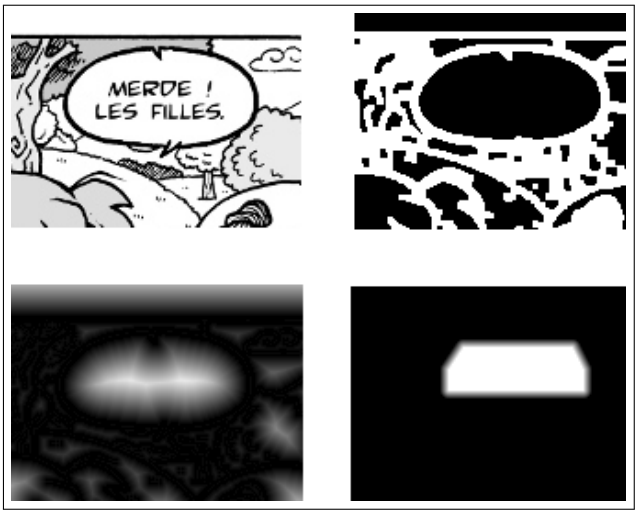


Figure 3.10: Example of original image (top left) and its corresponding non-text edge detection (top right), E_{ext} energy (bottom-left) and E_{text} energy (bottom-right). In the bottom part, white corresponds to high energy. Image credits: [38].

The energy E_{cont} forces the contour to be continuous by keeping points at equal distance, spreading them equally along the snake according to the average inter-point distance of the contour. It becomes small when the distance between consecutive points is close to the average (Equation 3.4).

$$E_{cont} = \alpha \left| \bar{d} - \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2} \right| \tag{3.4}$$

where $\bar{d}$ is the average distance between two consecutive points i and j of the snake and α is a weighting parameter.

The energy E_{curv} enforces smoothness and avoids oscillations of the snake by penalizing high contour curvatures (minimizing the second derivative). It becomes small when the angle between three consecutive points is close to zero (Equation 3.5).

$$E_{curv} = \beta \left((x_{i-1} - 2x_i + x_{i+1})^2 + (y_{i-1} - 2y_i + y_{i+1})^2 \right) \tag{3.5}$$

where i is a point of the snake and β a weighting parameter.

Text energy

The text energy E_{text} conveys domain specific knowledge about the relative locations of text areas and their associated balloon contours in this domain. It is necessary to consider the lack of explicit information in the cases of implicit balloons, where parts

of the outline are missing. The E_{text} energy aims at pushing the snake outwards toward the most likely balloon localization, given the position of the text area. This energy term has two effects. First, it acts collaboratively to the external energy, by moving the snake towards non-text edges (hopefully corresponding to the balloon outline). Second, in the case of implied contours where no explicit edge exists (the external energy term is not informative), E_{text} assists the algorithm to converge to an approximate contour position based on prior knowledge on the expected localization given the corresponding text area. We define the text energy term at localization i of the image as follows:

$$E_{text} = \begin{cases} \kappa \frac{N}{min_{j \in T} A(i,j)} & \text{if } A(i,j) > 0 \\ \kappa N & \text{else} \end{cases} \tag{3.6}$$

where j is a pixel in the text area T, N is an experimentally defined constant expressing the expected distance in pixel between the text area and the corresponding balloon boundary and κ is a weighting parameter that controls the contribution of E_{text} with respect to the other energy terms in Equation 3.2. Note when i is on the border of T, the distance $A(i,j)$ is equal zero and the energy becomes maximal as if it was inside T.

Proposed method

In this section we detail how to localize speech balloons using active contours based on the definitions given above. First we generate the static external energy map E_{ext} for the whole image and then for each text area we compute the E_{text} energy. The internal energy E_{int} is calculated for each point of the snake before each iteration. We iteratively examine each point of the snake in a clockwise fashion and move it within a neighbourhood region of size M in order to minimize Equation 3.2. This operation is repeated until no point moves in one turn (Algorithm 1). Detecting the implicit parts of the balloons is quite challenging as we aim at detecting something which is not in the image, by interpolating contextual information (other printed part of the balloon). We propose a two step approach to test if the information is present or not before trying to extract it precisely. First, we perform a coarse balloon contour position approximation (low resolution) using a quite rigid snake model. Then we fit the contour better to have a fine extraction of the balloon contour (high resolution) using the same algorithm and tuning the energy weighting parameters in a way to relax the snake (increase flexibility). The idea is to start a high resolution contour detection only when the active contour is positioned close to the real contour. Otherwise, if we start the balloon detection using a too flexible snake, it might be attracted by forces that do not come from a balloon edge before reaching the contour position (printed or suggested).

Active contour initialisation The active contour is initialized on the outline of the text paragraph region. This corresponds to the convex hull of all the text lines

Algorithm 1 Open balloon detection loop

```
compute E_ext energy
for each text area do
  compute E_text energy
  active contour initialization
  stop = False
  while stop = False do
    n = 0
    for each points of the snake do
      examine neighbourhood position energies
      if one position reduce the current energy then
        move point to this position
        n = n + 1
      end if
    end for
    if n = 0 then
      stop = True
    end if
  end while
end for
```

that are included in the balloon (Figure 3.11b). Note that the convex hull of the text area also corresponds to the E_{text} maximal value border (Figure 3.10).

The initial number of points impacts the way that the snake moves and the precision of the final detection. During the first low resolution localization step, we perform a spaced equipartition of the points (Figure 3.11c) to quickly localize the global shape avoiding unnecessary stops on image details. In the subsequent high resolution fitting stage, we add more intermediate points to fit the exact shape more precisely.

a) Group of text lines b) Text area convex hull c) Snake initialization

Figure 3.11: Active contour initialization based on text region convex hull. Image credits: [38].

Low resolution contour fitting Following a two stage process, we first aim to obtain a rough localization of the balloons by fitting a coarse contour using a few contour points during the initialization. The idea is to progressively push the snake away from the text area and towards the balloon boundary giving an increased weight to both the E_{ext} and E_{text} energy terms. If the balloon has an explicit boundary then E_{ext} will attract the snake to it. If there is no explicit contour close enough to attract the snake then the E_{text} term will push the snake to the suggested position of the balloon contour. Also the internal energies are important at this stage to maintain a certain rigidity of the snake. At the end of this step, we obtain a coarse description of the balloon contours which can be sufficient for localisation purposes but not for balloon type classification for instance (Figure 3.12).

Figure 3.12: Example of low resolution balloon contour detection (red line) for irregular closed (top) and smooth open (bottom) speech balloons. Image credits: [95, 129].

High resolution contour fitting Figure 3.12 shows that the global shape of the top balloon has already been retrieved, even for the implicit contour parts. The aim of the high resolution fitting process is to attract the snake to the balloon contour details where they are drawn and keep the implicit contour position found by the low resolution process elsewhere. To achieve a fine detection, we increase the resolution of the snake by adding new points between the original ones, changing the weighting parameters of the energy function and going through a second fitting process using the same Algorithm 1. At this stage, we relax the E_{curv} energy to make the snake fit thinner parts of the image and we set E_{cont} strong enough to keep a regular inter-point distance all over the contour. Also, we reduce the E_{text} energy weight because at this step, the snake is already far from text and this term is not informative any more. This new configuration allows the snake to fit more precisely to the balloon contour as shown in Figure 3.13 (to compare with Figure 3.12).

Figure 3.13: Examples of high resolution contour detection (red line) for closed (left) and open (right) balloons.

3.4 From balloon to tail

Tails are part of the balloon contour, they make the visual relation between the two most important elements in comics: balloons and comic characters (Section 2.2.2). Speech balloons indicate dialogue using tails pointing at their respective speakers [154]. From our knowledge, detection has not been studied before even-through it is key information for comics understanding. They are represented by a discontinuity on balloon contour (Figure 3.14).

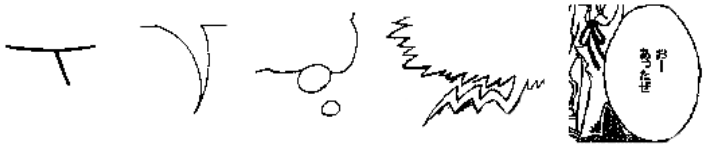


Figure 3.14: Examples of type of speech balloon tails. From left to right: stroke, comma, circle, zigzag, absent.

The first interesting information to extract is: where the tail is *connected* to the balloon and the position of its tip (extremity). From these two pieces of information we can compute a first direction and try to find the character to which it is related to in a post processing. Here comes the main difficulties, the tail may change direction several times from its origin to the tip, only the last part of the tail indicates the direction of the speaking character. Moreover, depending on the method used for extracting balloons, the tail tip position can be predicted in the background region or on the balloon boundary (more accurate) which may give a different direction at the end (Figure 3.15).

We propose to detect the tail position and direction based on the balloon contour analysis as tails are part of the balloon contour. We limit this study to the tail types that can be considered as an extension of the inside region (background) of the speech balloon namely types “comma”, “zigzag” and “absent” in Figure 3.14, because they can all be extracted from the segmentation of the balloon background. This limitation covers most of the speech balloons, nevertheless, other types such as “stroke” and “circle” in Figure 3.14, require different approaches for speech balloon

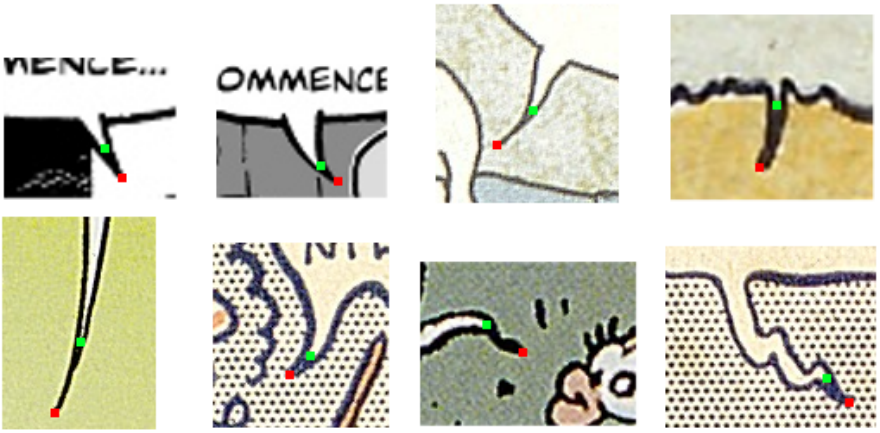


Figure 3.15: Tail tip position differences considering the balloon background region or the balloon boundary. In each vignette, the green and red points represent the tail tip position from the balloon background or boundary point of view respectively.

extraction (e.g. blob detection, contour boundary analysis).

The tail can be decomposed into four elements: origin, path, tip and pointed direction. We call "the origin" the virtual point where the tail is plugged to the balloon. The path is the symbolic line that represents the trajectory of the tail from the origin to its tip, we did not study this information in this work. In contrast, the tail tip position is an important information because it is the anchor of the tail direction which indicate, when combined, the location of the speaking characters. We propose a method based on convexity defects to detect the tail tip position and direction of "comma", "zigzag" and "absent" (no tail) tail types. For each tail tip position detection, we compute a confidence value in order to be able to detect "absent" tails when the confidence is too low. The directional angle of each tail is named with one direction corresponding to one of the eight cones of $\pi/4$ radians from the unit circle in order to facilitate further processing and evaluation (North, North Est, Est, South Est, South, South West, West and North West).

We observed that the contour of a speech balloon is mainly convex except in the region where the tail is plugged to the balloon (origin) which produces the highest convexity defects. A convexity defect is defined by a triangle from one segment of the balloon convex hull to the farthest point on the balloon contour (Figure 3.16). The set $F = \{f_0, f_1, ...f_n\}$ represents farthest points from the corresponding hull segments $S = \{s_0, s_1, ..., s_n\}$ where n is the number of hull segments. We define the top two farthest points f_a and f_b corresponding to the two farthest points of the convex hull segments s_a and s_b, as the coordinates of the tail origin (Figure 3.16).

We assume that the tail tip corresponds to one of the vertices $V = \{v_0, v_1, ...v_n\}$ of the convex hull (Figure 3.16), which is the case for most of the speech balloons

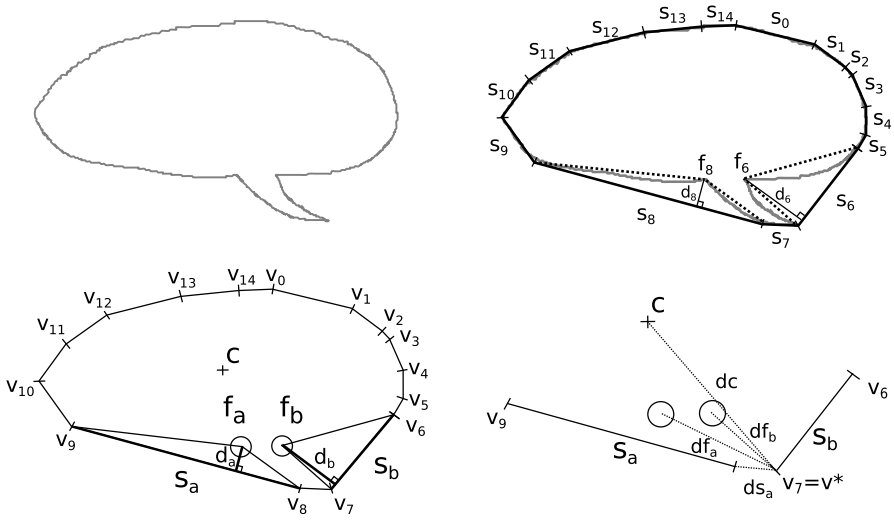


Figure 3.16: Convex hull and convexity defects of a speech balloon (top part). Bottom-left figure represents the two biggest convexity defects that defines the tail origin f_a and f_b. Bottom-right figure represents the distances related to the vertex v_7 to illustrate the variables of Equation 3.7. In this case $ds_b = 0$ because v_7 coincides with an end of segment s_b (distance equal to zero).

except when the tail is partially surrounded by the balloon (in a recess). We did not consider such cases because they are uncommon and require a different approach. The optimal vertex is computed by comparing five features that corresponds to the Euclidean distances:

- ds_a the distance to the segment s_a
- ds_b the distance to the segment s_b
- dc the distance to the centre of mass c of the balloon
- df_a the distance to the tail origin f_a
- df_b the distance to the tail origin f_b

This can be formulated as:

$$v* = argmax\big(max(dc + df_a + df_b) + min(ds_a + ds_b)\big) \tag{3.7}$$

where $v*$ is the optimal vertex from the set of vertex V. In Figure 3.16, the optimal vertex is v_7.

From our knowledge, all closed or implicit balloon extractors proposed so far, consider the inside balloon region and ignore the balloon contour (black stroke) region.

This is acceptable for localising the tail on the balloon contour (angular position) but it is not accurate enough to detect the tail tip in the sense of human understanding. In fact, the tail tip is often located at the extremity of black stroke surrounding the inside balloon region (Figure 3.15).

We propose to detect the external border line of the balloon contour in the region of the optimal convex hull vertex $v*$ presented above and define its extremity as the tail tip position. Image pre-processing and contour detection are performed in the same way as presented Section 3.3.1. From all the detected contours, we select the contour O that has the closest point o_i from $v*$ (Euclidean distance). Then we run through all the points of the sequence of points that define the contour $O = \{o_0, o_1, ..., o_m\}$ and define as tail tip position $o*$ the one at the maximum Euclidean distance to the tail origin. The tail origin is a virtual point midway between f_a and f_b (Figure 3.17).

The tail direction is given by the last part of the tail which corresponds to the vector $\overrightarrow{v*o*}$. Note that starting from $v*$ is more accurate than the tail origin point because the last segment of the tail is oriented towards the speaker while the tail origin is just "getting out" of the balloon.

Figure 3.17: Tail tip position and direction detection. From left to right, the optimal vertex $v*$ in green and the local contour detection in blue, the closest point to $v*$, o_i in red and its first stop at the maximal distance with the tail origin (dark green point) $o*$. The last vignette on the right hand side represents the tail direction vector $\overrightarrow{v*, o*}$ by a white arrow.

We also propose to compute the confidence value of the tail position prediction C_{tail}, related to the mean of the depths d_a and d_b over the mean balloon size $meanBalloonSize$ (Formula 3.8). The confidence value is minimal when the mean of the two convexity defects (used for tail position detection) is very small compared to the balloon sizes in the current image. It is maximal when equal or higher than the mean balloon size.

$$C_{tail} = \frac{(d_a + d_b)/2}{meanBalloonSize} \tag{3.8}$$

where d_a and d_b are the depth of the two biggest convexity defects in number of pixels (Figure 3.16) and $meanBalloonSize$ is defined in equation 3.9.

$$meanBalloonSize = \frac{\sum_{i=0}^{n} W_{b_i} + \sum_{i=0}^{n} H_{b_i}}{n * 2} \tag{3.9}$$

where W_{b_i} and H_{b_i} correspond to the width and the height of a speech balloon i and n the number of speech balloons in the image.

3.5 From tail to comic character

As presented Section 2.2.4, unsupervised comic character detection is still an unsolved issue. In fact, it faces two difficulties. The first one is the variety of comic styles that directly impacts the diversity of comic characters as they are the main elements of the stories (Figure 2.4 and Appendix C.1). The first challenge is to propose a generic approach able to cover such variety. The second challenge is due to the variations of position, size, pose, scale, occluded parts between each instance of a given comic character (Figure 2.8).

Here we propose an approach that bypasses both issues by using already extracted information in order to predict the locations of the comic characters. As we know which balloons are in which panel (total inclusion), we can estimate from the position and direction of the tail, which part of the panel may contain the speaking character. This approach is limited to speaking characters but can be post processed in order to find non-speaking characters given the speaking ones as examples (Section 5.4.2).

As mentioned in the previous section, the tail direction is quantized in eight cones of $\pi/4$ radius for simplification purposes so we can consider that, in a rectangular-shaped panel (or its bounding box), the tail is either pointing towards a corner or towards a border of the panel. We define the ROI for the character as a squared region beside the speech balloon. The maximum width w_{max} and height h_{max} of the ROI are equal to the mean widths and heights of all the balloons in the image (Equation 3.9).

The position of the ROI around the speech balloon is defined by two opposite points of a square $A_{x,y}$ and $B_{x,y}$ according to formula 3.10 which is sometimes constrained by a panel border that is why the term min appears (Figure 3.18).

$$\begin{array}{rcll} A_x & = & v_x^* - min(Pi_x - v_x^*, w_{max}) & *O_x \\ A_y & = & v_y^* - min(Pi_y - v_y^*, h_{max}) & *O_y \\ B_x & = & [A_x + min(Pi_x - A_x, w_{max})] & *O_x \\ B_y & = & [A_y + min(Pi_y - A_y, h_{max})] & *O_y \end{array} \tag{3.10}$$

where v^* is the coordinates of the tail tip (Section 3.4), Pi the coordinates of one of the four corners of the panel bounding box $P = \{P0, P1, P2, P3\}$ and O a shift value. The shift value allows an horizontal and vertical translation of the ROI position relatively to the tail tip direction (Figure 3.18). The shift O is quantized in heigh values, according to the tail direction quantization (Section 3.4), for horizontal and

vertical axis. The panel's corner P_i is chosen between two opposite corners in order to always give a positive value in the computation of the ROI position (Table 3.1).

Table 3.1: Values of the horizontal and vertical shifts and panel's corner selection according to the eight directions of the tail.

	N	NE	E	SE	S	SW	W	NW
O_x	0.5	0.75	1.0	0.75	0.5	-0.75	-0.5	0.75
O_y	-1	-0.75	-0.5	0.75	1.0	0.75	-0.5	-0.75
Pi	P1	P1	P1	P1	P3	P3	P3	P3

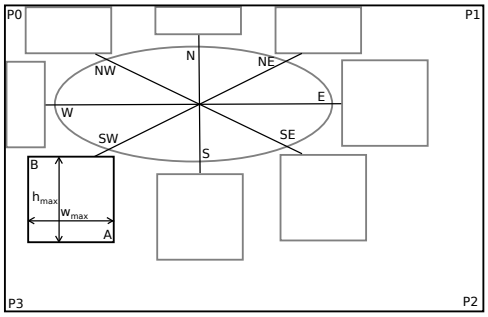


Figure 3.18: Illustration of a panel with four corners $P0, P1, P2, P3$ and the different ROI position and size that can be defined for each of the eight directions of the tail.

3.6 Conclusions

In this section we have presented how to benefit from simple elements to retrieve more complex ones in a comics image by following the relations between elements. This approach is similar to the reader's reading behaviour opening a new comic book for the first time: first looking for traditional elements such as the panels, text and speech balloons and then, reading their content and retrieving the links with the comic characters and other graphics in the panel. This first approach is driven by the image content which means that each newly extracted element allows other related elements to be retrieved. It is particularly useful for ensuring a certain consistency between all the extracted elements but in return, the dependency of the processes propagates extraction errors. For instance, if a panel is not extracted then its content will not be processed and potential balloons and comic characters will be missed. The experiments, presented in Chapter 6, evaluate and compare this approach to methods from the literature and other approaches that will be presented in the two next chapters.

In the next chapter a non-sequential approach will be presented in order to extract comics image content using independent processing and so, avoid error propagation

which is the main drawback of applying several consecutive processing.

Chapter 4

Independent information extraction

In this chapter we propose an independent extraction of each element contained by comic book images in order to make the image interpretation more flexible. This approach tries to fill the weaknesses of the sequential approach presented in the previous chapter by making each specific content extraction algorithm independent from each other in order to avoiding error propagation. New approaches for extracting panels, text, balloons and comic characters are presented successively but they can be used in any order or in simultaneously as they are totally independent from each other. Reaching such level of independence requires embedding more domain knowledge in each processing algorithm in order to give then the ability to autonomously process an image without relying on previous processing.

4.1 Introduction

Image content retrieval methods are useful for the community and the society when they can easily be applied on various types of images sharing similar properties, independently from other pre or post processing. Another important aspect is to propose methods with a minimum number of parameters or at least that do not make the method too specific but generic in order to be usable for a maximum of applications. The first approach presented in Chapter 3 is very specific to images from comic books because each extraction of element requires one or several pre-processing (other element extractions). In this chapter we present several contributions to extract elements independently from each other. Each contribution is then easily reusable for specific needs with a lower processing time (e.g. balloon locator independent from panel positions, comic character spotting without needing panel and text positions). The independence between the processing is also useful in a content of fully automated system to allow local optimisation without influencing the whole sequence and output. Independent panel, text and balloon extraction methods are proposed, additionally, text recognition and balloon classification complete the amount of information that

is retrieved from the image content, towards a semantic description. The problem of comic character retrieval is simplified here as a character spotting giving an example as query for the reason presented Section 1.3.

4.2 Panel extraction

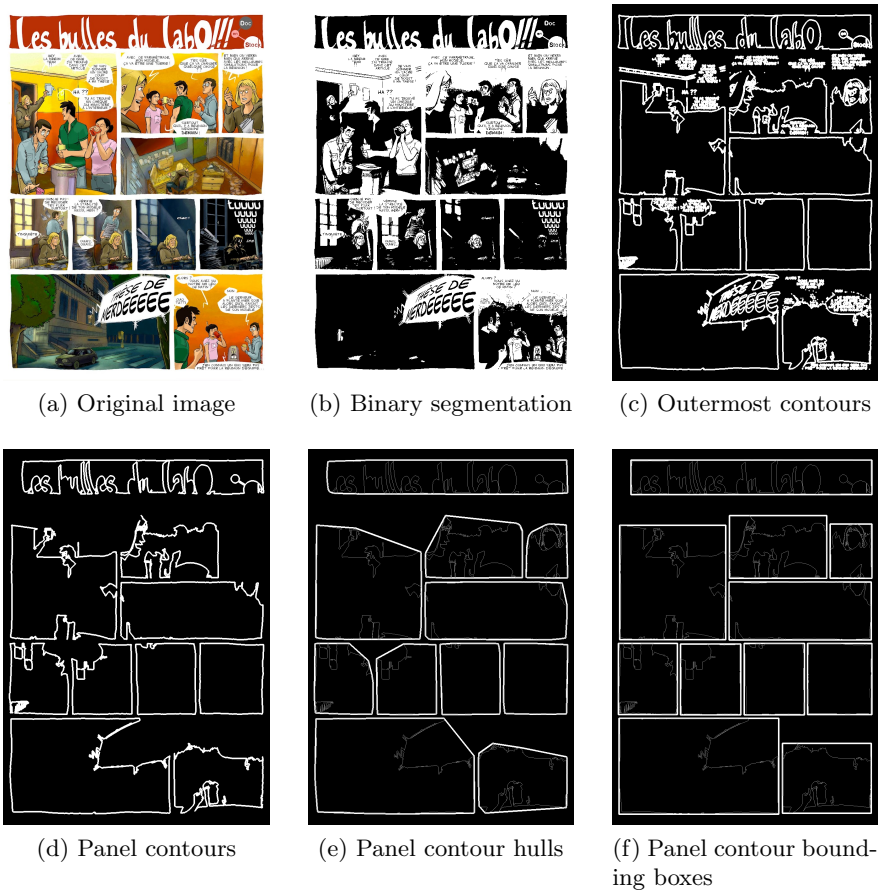
The panel extraction method presented in the previous chapter is based on connected-component clustering (sequential approach Section 3.2). This is particularly useful for simultaneous extractions of panel and text while removing noise regions at the same time. The main issue comes from the clustering algorithm that uses a fixed number of clusters and assumes that the image contains connected-components from panels, text and noises (related to each other). This is not always the case, sometimes the amount of text or noise is not significant and confuses the clustering. We propose to overcome this limitation by extracting panels independently from other elements. In the first approach, the connected-components that are labelled as panel by the clustering algorithm are then pruned based on a topological filter (Section 3.2). Here we base our independent approach on this topological filter by considering as panels only the connected-components that are not included in any other connected-component (no parents). In Figure 3.1b, we show that outermost or external contours actually corresponds to panels. We base this method on this criterion, considering the panels to be the outermost contours. This is especially true for general comics using gutter (white space) between panels. Making the outermost contour as panel can be seen as a binary classification method deriving from a more generic concept. The general concept would attribute a confidence value to each contour in the image, inversely promotional to their percentage of inclusion in other contours. Non-included contour (outermost) would have the maximum confidence and fully included ones the lowest confidence in being a panel.

Panel contours (border line) are usually dark and similar for all the panels of a comics page. Sometimes panel contours are implicit as shown Figure 4.1a (considered as a difficult case). In this case, the reader distinguishes them by comparing the difference of colour between the panel content and the page background. We use a global binary segmentation method (Appendix A.1.4) to separate the page background (expected to be clear) from its content (dark elements such as black strokes), considered as foreground here (Figure 4.1b). A binary inversion should be performed when the image background is darker that the content of the panels. We do not use a local segmentation approach here because it has higher chances to split the panel border and over-segment its content.

Then, a connected-component labelling algorithm (Appendix B.1) is used to extract outermost contours from the binary image, (Figure 4.1c). Note, we filter out contours below a minimum area relative to the image size $minAreaFactor$ in order to avoid considering isolated elements (e.g text, logo, page number) as panel in a later stage (Figure 4.1d)

Finally, the convex hull or the bounding box of the panel contours can be com-

puted in order to recover from discontinuous contour detection (Figure 4.1e and 4.1f). In this example, some of the panel borders have low contrast compared to the page background which makes difficult their extraction. Such issue can be post processed using line detection and layout analysis.



(a) Original image (b) Binary segmentation (c) Outermost contours

(d) Panel contours (e) Panel contour hulls (f) Panel contour bounding boxes

Figure 4.1: Panel extraction process. Image credits: [107].

4.3 Text localisation and recognition

In this section we present a method for automatic text localization in scanned comic books, an essential step towards an automatic comics understanding. Our first intention is to localize multi-oriented text regions, independently from script and language, and separate them from the graphical content (text/graphic separation). In document analysis systems, text recognition quality often relies on text localization. This is particularly the case for comics which have a complex background. We present how to

take advantage of Optical Character Recognition systems (OCR) in order to improve text localisation (conventional paradigm of recognition after segmentation [206]).

4.3.1 Introduction

In comic book documents, text is generally mixed with graphical contents, what produces a lot of false detections if we submit the whole image directly to an OCR system, usually coupled with a layer analysis pre-processing to avoid trying to recognize text in graphical regions (e.g images, tables and formulas). The mixture of typewritten, handwritten-like and handwritten text makes the task particularly difficult for comics. Moreover, text presents a lot of variations (e.g. stroke, orientation, colour, size) that can cause many issues at both the localization and recognition levels. Text localization aims to provide text-only (or graphics-free) images to a text recognition system in order to improve its accuracy and reduce computing time. Knowing the positions of other elements in the image is a good clue for predicting text position because they are related to speech balloons (e.g. dialogue, thought), characters (e.g. graphic sound, illustration), panels (e.g. captions) or the page itself (e.g. page number, title, author name). This approach is discussed Chapter 5.

Few works concern text extraction in comics analysis literature (Section 2.2.3) and most of them rely on speech balloon regions which make them dependent on other processing quality. Here we want to avoid such dependency between processes. From our knowledge, only Li [101] proposed an independent and unsupervised speech text localization. It is a two-step analysis that first generates text lines using "a set of very rigorous rules" to test each pair of connected-components and decide if they are candidate text regions or not. From the first-round generated text lines, the method automatically learns a font set which is propagated to all text lines in the second-round and filters out non font-matched candidate regions. The font set is generated from the distribution of heights and widths of the connected components that compose the candidate text line (composed by multi-segment characters). The method is promising but there are nine rules in total that must be met to achieve text line extraction with several heuristics that are not discussed. The experiments have been performed on 1000 images from 10 countries (balanced) but they are unfortunately not publicly available to assess their diversity and the genericity of this approach.

The genericity is always a challenge is document image analysis systems, especially because it requires an additional effort for researchers to validate their methods on several other application domains (Section 2.1). Moreover, comic book images differ from classical documents in that they comprise complex backgrounds of a graphical nature. They belong to the class of non-structured documents meaning there is no regular structure present for the prediction of text locations and no layout analysis method applicable if we consider all the diversity of comics. Comics being unstructured graphical documents, combine the difficulties of both domains, making the task of text localization especially challenging. To solve the particular problems which are provoked by the combination of complex background and unstructured documents, we propose a new text localization method.

Our contributions come at different levels. First, an adapted binary segmentation method is presented followed by a text/graphic separation algorithm based on the local contrast and neighbourhood similarities and finally, text recognition is used to verify the presence of textual information in the result. The advantage of the proposed processing sequence compared to other methods from the literature is its ability to retrieve multi-script and multi-oriented text with complex background in an unsupervised way.

4.3.2 Bi-level segmentation

Segmentation is a crucial step in many text localization methods. Comic speech text is made of strokes, generally black on white, which we would like to isolate from the rest of the image content (background). A perfect text segmentation would result to individual text letters represented by single connected components (Appendix B.1). The complex background of comic documents complicates this step. Separating text from other elements in once is quite hard since we have no information about its position (independent approach) and it is similar to many other strokes in the image. Therefore, we propose first to separate bright from dark regions and then classify text and non-text regions using size and alignment properties only. This approach is unsupervised and does not require any training step. Bright/dark region separation can be done using a bi-level segmentation algorithm. Several approaches exist and their performances rely on the input data distribution. If the input data are composed by only black or white pixels, any method will work but this is rarely the case. The best input would be a highly contrasted image where the features of interest are concentrated in one of the extrema. We convert the 3 channel colour image (RGB) into a single channel image (grey-level or grey-scale image). There are several manners to create a grey-scale image from a colour image. The first approach is to combine each pixel value from each colour channel with different weighting coefficients corresponding to the measured intensity perception of typical trichromat humans[1], in order to produce a single value per pixel. Another approach is to use only one channel but as the text we are interested in is not in red, green or blue (RGB colour space) it is irrelevant in our case. Nevertheless, RGB colour space can easily be converted to a different colour space such as HSV or HSL that encodes colours as Hue, Saturation, Value or Luminance. The luminance channel is particularly appropriate for bright and dark region separation [67].

Both colour layer combination and separation are experimented Section 6.4.3. From the grey-scale image, we determine the threshold value that will best divide the grey value distribution into two clusters. The ideal would be a cluster containing pixels from text regions (foreground) and another one with the rest of the content (background). In practise, the foreground cluster often contains text plus other similar grey value regions (e.g. black strokes from contours) and the second one other pixels with different grey values (e.g. bright graphic regions).

Fixing a threshold value like Arai's method [8] works only for few comics types

[1] `http://en.wikipedia.org/wiki/Grayscale`

with a uniform text background intensity (Figure 4.2). This phenomenon is intrinsic to the nature of comics, tightly linked to the design process. Indeed, comics generally contain textured areas and loosely connected strokes that give rise to merged components at different thresholds. This is intensified by the digitization and image compression process that add further noise to the page (Section 2.1). Instead of using a fixed threshold, we use the well-known Otsu's threshold selection method that has been extensively used since decades for this purpose but other methods would be appropriate as well (Appendix A.1.4). Otsu's threshold is computed for each image and applied to the whole image assuming that the image has not been degraded locally. In case of such local degradation (e.g. holes, crossed text, partial erasure, highlighting, tearing paper, stain), a local threshold selection method should be used (Appendix A.1.4).

As an alternative segmentation method, we experimented text region extraction using the Maximally Stable Extremal Region (MSER) algorithm [113]. This algorithm was better than the proposed method only for text region of type onomatopoeia but produced an excess of false detections as in general a lot of the graphical elements are equally stable as the text (flat regions).

4.3.3 Text/graphics separation

After the binary segmentation step we apply the connected-component (CC) labelling method (Appendix B.1) to extract the black strokes as many previous work in the literature (Section 2.2.3). At this stage, the extracted components may correspond to part of drawing, single letters or a sequence of letters if some are connected (Figure 4.3). The objective of this section is to separate the last two categories from the first one. Note that connected letters will not affect our results (if they are part of the same word) because the expected output is at text line level.

Due to the wide variety of text usage in comic albums, text and graphics separation is not obvious, the method should be size, position, rotation, scale and contrast invariant. In addition, it should be robust to text-less pages which may appear randomly within some albums. From all the components extracted by the CC algorithm, we select only the ones corresponding to potential textual elements based on contrast ratio measure.

We compare the standard deviation of the pixel grey level distribution of each CC bounding box (Figure 4.4) with the contrast ratio of the page to make a low/high local contrast CC separation, similarly to [101] using the Mahalanobis distance. Mahalanobis distance measures how many standard deviations away a point is from the mean of the distribution [42, 110]. We call this distance d_m and assume a Gaussian distribution of the grey level pixels but other distribution are under study. The contrast ratio of the page is the absolute difference between the minimum and the maximum grey-levels found on the page.

Figure 4.2: Different threshold selection methods applied on the luminance channel of several comics types. First row shows the original image of a selection of comic book parts. Second row shows the corresponding luminance channel of HSL colour space in a 8 bits grey-scale image. The third row represents the bi-level segmentation of the grey-scale image with a fixed threshold at 250 (as Arai [8]) and the fourth row shows the corresponding results using the automatic Otsu's threshold selection method. Image credits: [38, 74, 115, 134].

Figure 4.3: Example of connection between connected components. On the left, a word with six well separated letters, on the right a word detected as two pairs of attached letters because of the handwriting variability and segmentation process.

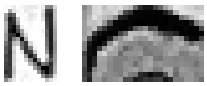


Figure 4.4: Example of mean (μ) and standard deviation (σ) values of two CC bounding box regions. For the letter "N", $\mu, \sigma = [167, 84]$ and for the eyebrow $\mu, \sigma = [110, 57]$.

4.3.4 Text line generation

As already mentioned Section 3.2, the gap between text component (e.g. letter or attached letters, word) and text lines can vary significantly. Here we use the same method as presented in Section 3.2 to group text component into text lines considering only the text height (or width for vertical text).

4.3.5 Text recognition

Automatic text recognition has been studied for many years, today OCR systems are very efficient for standard fonts but not optimized for comics fonts which are mostly handwritten. The principal interest of text recognition is the ability of searching for images by using keywords which opens up many applications (Section 2.2.3). Text recognition applied to comics is really challenging because it includes most of the difficulties from text recognition in the document analysis domain if we consider all the types of text that compose the comics. The difficulties are due to typewritten, handwritten or free-form text on uniform or complex background including image noise, text deformation and overlapping (Figure 2.7). Note that the text is mainly written in upper case for Latin scripts in comics which reduces the complexity of recognition.

Text recognition for comics requires extensive work to be addressed properly, which is out of the scope of this thesis work. Even though standard OCR recognition rate is not acceptable yet for full-text based application in comics, OCR can help at separating text/non-text regions.

We pass each text line candidate through an OCR system to validate its content. The validation criterion is that the OCR recognizes at least one alphabetical symbol within the candidate region, otherwise it is rejected. Text lines are composed by 12 characters on average which give a good chance for OCR to recognize at least one of them (Figure 6.1).

4.4 Balloon extraction and classification

In this section we present a method for pixel level balloon extraction and for balloon classification. Here, balloon extraction does not require previous text extraction un-

like the method presented Section 3.3.1,instead it uses a careful analysis of the region content (e.g. similarities, spacial organisation) to predict which are most likely balloon regions. Balloon contour and shape information differences are presented and a contour-based classification method is detailed for "smooth", "wavy" and "spiky" classes that may correspond to "dialogue", "thought" and "exclamation" in a higher level context (Chapter 5).

4.4.1 Balloon segmentation

From the reviewed methods Section 2.2.2, we can see that the text position often helps in the localization and then segmentation of speech balloons. The issue in relying on text extraction is that we become dependent on its performance. Text extraction in comics images is still a challenging problem (Section 4.3), it is a very hard assumption to assume that text extraction will work perfectly. Instead of relying on text detection we propose to keep both processes independent in order to avoid propagation of errors. In a similar way, Arai's method [8] does not use text position but connected component filtering based on heuristics (fixed thresholds for binary image segmentation, blob size, number of white pixels, number of contained straight lines and width to length ratio). We propose a more adaptive method in order to handle various comics styles (e.g. colour, B&W), image formats (e.g. A4, A5, portrait, landscape) and digitization variations (e.g. scanning device, camera capture).

The use of fixed thresholds for the segmentation process such as [8], is a strong assumption about the image background nature. It only works using an empirical approach for choosing the threshold value and is appropriate for comics of the same style and digitized under the same light condition. We propose to relax the constraint by using a binary segmentation method that automatically selects the optimal threshold value. Assuming that all the parts of the scanned image have been digitized under the same condition allows us to apply the same threshold for the entire image. The well-known Otsu's threshold selection method [137] have been proposed to find the best threshold that separate clear from dark regions. Balloon components are composed by a clear inside region surrounded by a black stoke which corresponds to what Otsu's method have been designed to. We use Otsu's method to define the optimal threshold and apply it on the whole image in order to make a global binary segmentation of the grey-scale image (Appendix A.1.4). Other optimal threshold selection methods from the literature can also be used as this is not a crucial point of the method. Note that bi-level segmentation (binarisation) creates a lot of small connected-components on textured regions; they are not disturbing the process as we are interested in balloons which are quite big regions. We prune small connected-components using the clustering algorithm presented in Section 3.2 (ignores the connected-components that are part of the cluster with the lowest average height). We extract connected components from the binary image and analyse their parent-child relationship to make a first selection of candidate components and then analyse the spatial organisation of the set of children $CH = \{ch_1, ch_2, ..., ch_n\}$ to compute a confidence value for each parent $P = \{p_1, p_2, ..., p_m\}$ that serves as final filtering criterion (Figure 4.5).

Figure 4.5: Balloon extraction process. Original image, binary segmentation, contour detection (connected components) and result mask from left to right. Image credits: [124].

We assume that a balloon has a minimum number of children *minNbChildren* which are horizontally or vertically aligned and centred inside the balloon region (property of text). The orientation of alignment is set manually according to the language of the comics (e.g. vertical for Japanese, horizontal for English). The percentage of alignment *align* is computed for each child ch_i considering $CHA(ch_i)$ the subset of children that are aligned to ch_i (Equation 4.1).

$$align(ch_i) = \frac{|CHA(ch_i)|}{n} \tag{4.1}$$

where $|CHA(ch_i)|$ is the number of aligned children and n the total number of children.

For instance, if we consider two children E and F, they are considered as aligned (vertically) if the following conditions are verified $centroidF_x \in [minE_x, maxE_x]$ and $centroidE_x \in [minF_x, maxF_x]$ where min_x and max_x are the left and right limits on the horizontal axis. The child F is also considered to be aligned (horizontally) to E by changing x to y which become the top and bottom bounding box limits.

We use the difference of alignment between the children components as a clue about the probability of a parent component to be a balloon. This is similar to the text dependent balloon extraction method presented Section 3.2. The difference is that here we have no idea about the nature of the children (not detected as text before). This independent balloon extraction method can be seen as an "intelligent" balloon extractor that embeds a part of the domain knowledge related to text (balloons are expected to contain text). The difference of alignment is computed for the horizontal (d_x) and the vertical (d_y) axis from the Euclidean distance between the parent and children hull centroids (Figure 4.6).

Both differences d_x and d_y are normalized between zero and one as a percentage of the balloon (parent) width and height respectively. The average of the alignments, $\overline{align}$, d_x and d_y gives a confidence value for each candidate of P (Formula 4.2).

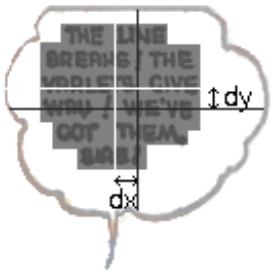


Figure 4.6: Horizontal and vertical alignment measurement between the centroids of the balloon (dark cross) and the children hull (clear cross). The children hull is the grey region included in the balloon.

$$C_{balloon} = \frac{1}{3} * \left(\overline{align} + d_x + d_y\right) \tag{4.2}$$

The confidence value is used for balloon and non-balloon decision (Section 6.6).

Tail detection is not treated as an independent extraction because it is part of balloons and never appears without being related to a regular or implicit balloon. Please refer to Section 3.4 for more detail about tail detection.

4.4.2 Balloon classification

Speech balloon classification is a real challenge if we consider a lot of different comic styles. First, each author has his own representation of the speech sounds. Second, the reader interpretation is subjective (each balloon style is interpreted in comparison to the already read balloon styles). For those reasons, it is really hard to directly define the purpose (e.g. dialogue, thought and exclamation) of a speech balloon without knowing its context in the story. In fact, the context is defined by both graphic (e.g. other strokes, protagonists personality) and textual (e.g. vocabulary, punctuation) elements. For instance, Figure 4.7 shows six different speech balloon comics styles where we can note the difficulty to assign one class for each balloon without knowing other balloon shapes in the image and reading the contained text.

As introduced in Section 2.2.2, the discriminant information for speech balloon classification is not provided by the global shape but by the contour variations. There are no previous approaches in the literature, we briefly reviewed state of the art approaches concerning shape and contour classification (which are the closest fields of research) but none of them matches our specific issue (Section 2.2.2). Therefore, we propose a new approach that separates contour from shape information, analyse the contour variations and then, describes and classifies the balloon contours into three classes (see “smooth”, “wavy” and “spiky” contour types in Figure 2.5). Note that this approach requires balloons to be extracted at pixel-level such as proposed in Sections 3.3.1 and 4.4.1.

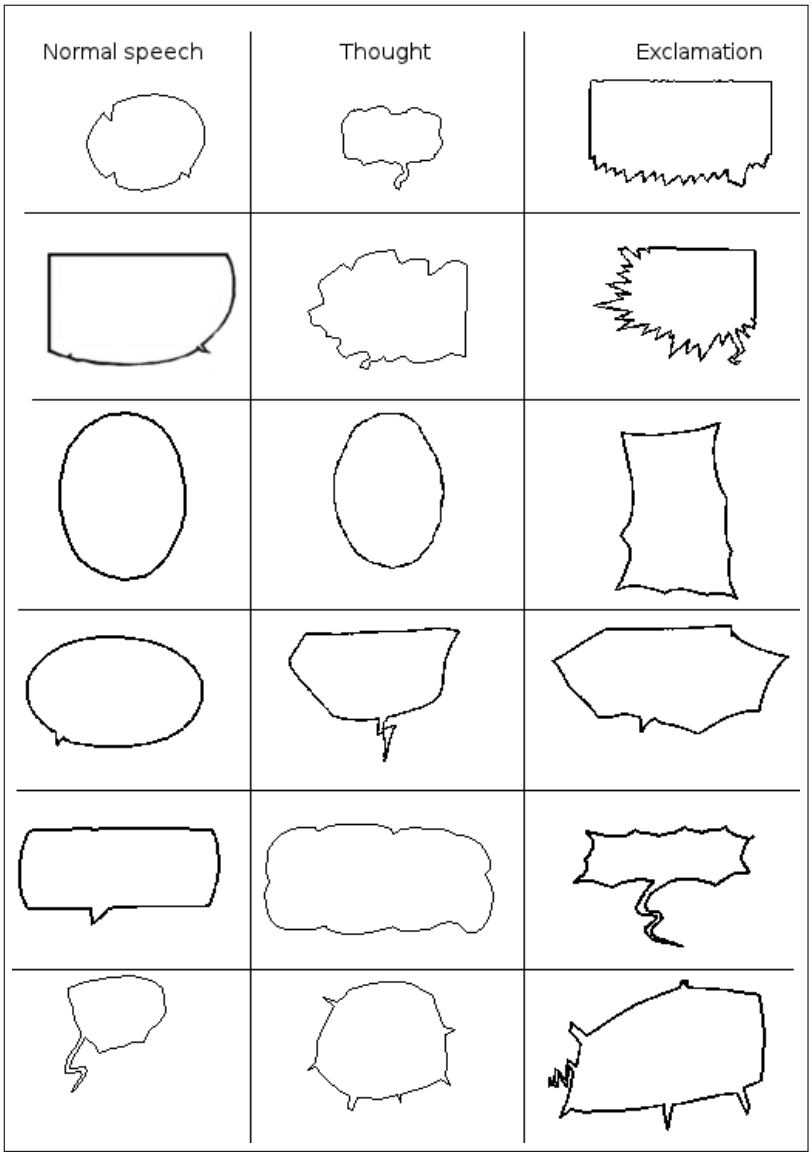


Figure 4.7: Examples of speech balloons from 6 comics albums (rows) that represent 3 different expressions (columns).

Shape/contour separation

Balloon shapes and contours contain different types of information. The overall shape (e.g. rectangle, square, circle and oval) does not provide a lot of information about how the text is spoken, it is more related to the style of the comics and the structure

of the panel. However, the contour variation describes the speech tone information according to the different patterns which compose the contour (e.g. straight lines, curves, peaks, half circles).

First, we propose to represent the speech balloon as a time series (one dimensional signal over 360°) which corresponds to the distance in number of pixels between the balloon barycentre and the contour points.

Second, we perform a shape/contour separation to be able to analyse only the contour variation. Note that the consideration or not of tail region has an impact on the final classification (Section 6.7). We approximate the global shape s by smoothing the original signal o using a sliding window of size M and subtracting the result from the original signal to preserve only the contour (high frequency) information c independently from the shape: $c = o - s$. The smoothed contour s is a centred local average of o (Formula 4.3). Examples are given Figure 4.8.

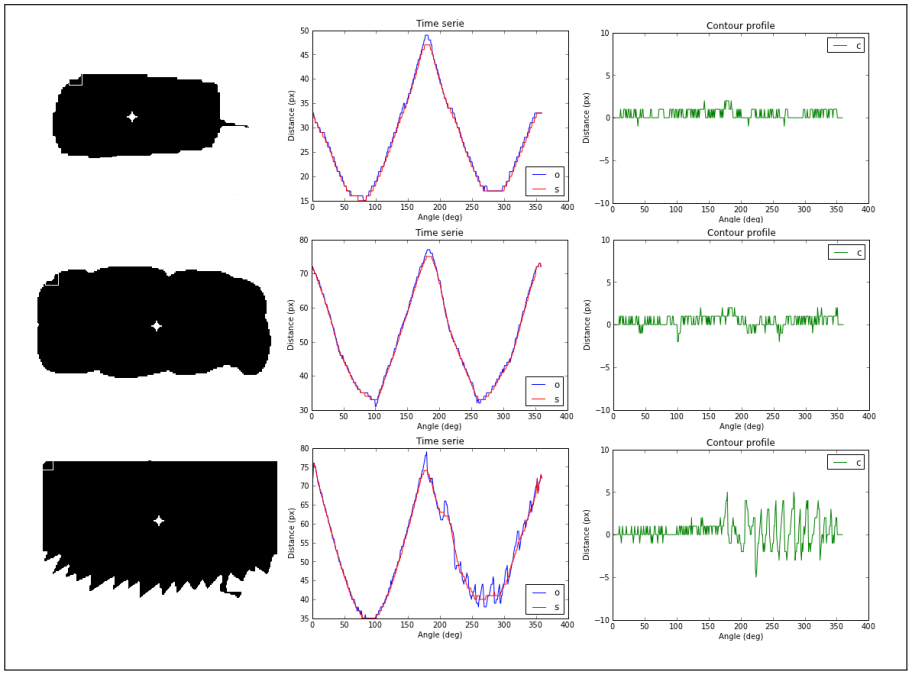


Figure 4.8: Balloon detection, time series representation and shape/contour separation. Left column represents balloon detection with the barycentre (white star) and the starting point (top left half square). Middle column represents the distance between the barycentre and the balloon boundary over 360° (clockwise) where the original signal o is in blue and the smoothed signal s in red. Right column is the difference between s and o.

$$s(o) = \frac{1}{M} * \sum_{-M/2}^{+M/2} o \tag{4.3}$$

Contour description

Contour description aims to encode discriminant contour variations in order to facilitate the classification step. The features used for the description must be as discriminant as possible but also independent to the page format and definition as introduced in Section 4.4.1.

The contour variations can be computed from different features, the main difficulty is to be invariant to the comics drawing diversity. In this study, we describe the contour signal c by using a variance-based two dimensional descriptor, more precisely the standard deviation σ. The variance or the standard deviation has the advantage to be simple and generic at describing any data distribution by measuring the data dispersion which is what we need if we aim at measuring the difference of dispersion between the original and the smoothed contours.

The first dimension aims to differentiate contours which have high variations from the others (e.g. "spiky" from "wavy" and "smooth" contour types in Figure 2.5). The second dimension aims to discriminate "wavy" and "smooth" types according to the standard deviation of the superior and inferior signal parts from the average $\bar{c}$. For instance, if a contour has a higher standard deviation in the below signal part than the above signal part (according to the average $\bar{c}$) then the contour has more peaks in the direction of its barycentre than the opposite, which is not a characteristic of the "smooth" but "wavy" type (half circles).

The first feature f_1 consists in measuring the standard deviation σ_c of signal c normalized by the local average of the original signal $\bar{o}$ in order to be independent to image format and definition variations (Formula 4.4).

$$f_1 = \frac{\sigma_c}{\bar{o}} \tag{4.4}$$

where f_1 is the feature one, σ_c the contour standard deviation normalized by $\bar{o}$ the average of the signal o.

In order to measure the second feature f_2, we split the contour signal c in two signals c_{pos} and c_{neg} which correspond to the signal parts that are strictly above and below the average $\bar{c}$ (Figure 4.9) and then we measure the standard deviation differences (Formula 4.5).

$$f_2 = \frac{\sigma_{neg} - \sigma_{pos}}{\bar{o}} \tag{4.5}$$

where f_2 is the feature two, σ_{pos} and σ_{neg} are the standard deviations of the signals c_{pos} and c_{neg} respectively. The difference is normalized by $\bar{o}$.

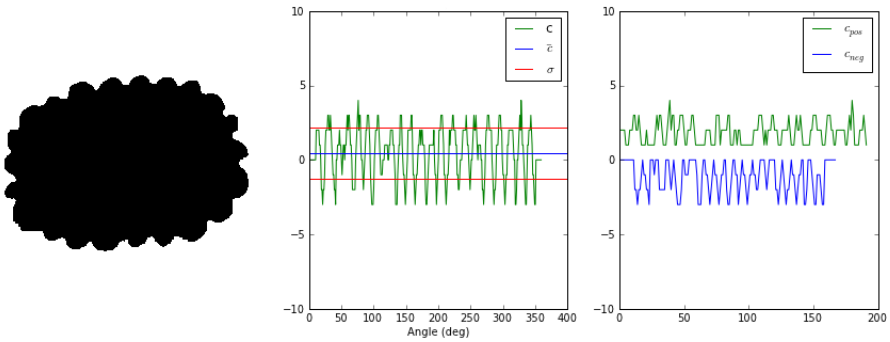


Figure 4.9: Contour signal decomposition. Left part is the concerned speech balloon, middle part is its contour profile and right part is the positive and negative decomposition according to the contour average $\bar{c}$. Both signal have not the same length because we ignore values that are equal to the average value $\bar{c}$ in order to not bias the standard deviation.

In order to be invariant to the image definition and scale, we normalize each feature of the descriptor by the original signal average $\bar{o}$ (average radius).

Contour classification

Contour classification is performed according to the selected contour descriptor (Section 4.4.2) by using a naive Bayesian Classifier [127]. This classifier has been chosen because it requires a small amount of training data. Let a label set $L = (l_1, l_2, l_3)$ represent the three contour classes "smooth", "wavy" and "spiky". Given a new unlabelled contour c described by its descriptor $D = (f_1, f_2)$, the naive Bayes approach assigns D to a class l_{MAP} as follows:

$$l_{MAP} = \text{argmax}_l P(l|D) = \text{argmax}_l \frac{P(l)P(D|l)}{P(D)} = \text{argmax}_l P(l)P(D|l) \tag{4.6}$$

where $P(l)$ is the *a priori* probability of class l and $P(D|l)$ is the conditional probability of a descriptor D given the class l. We approximate the probability density function as a Gaussian distribution for which we learn the parameters during the training step.

The results concerning balloon classification are detailed Section 6.7.

4.5 Comic character spotting

Human detection in computer vision field is an active field of research. Extending this to human-like drawings such as the main characters in comic book stories is not

trivial. The detection of the main comic characters is an essential step towards a fully automatic comic book understanding. This section presents a colour-based approach for comic character spotting using content-based drawing retrieval and colour palette descriptor.

In this section we detail how to localize comic characters in all pages of a given album. We aim at detecting the apparitions of comic characters (non-rigid object) with a minimum of user interaction. This is a content based image retrieval problem where the result highly depends on the quality of the given example, which should be noise free (no background, target only). A close up view of comics drawings shows that the most invariant information between texture, shape and colour is the colour information (Figure 2.4). In fact, they are no textures in many newer comics because they are designed and reproduced with colour flats (see comics design process Section 2.1), nevertheless, the digitization process sometimes creates textures (e.g dithering, compression). The shape information suffers from the different postures that can take the comic characters in each panel of the story (Figure 2.8). Hence, we believe that the colour information is the most robust information invariant to scale, object deformation, translation and rotation transformations which are very frequent in comics. The typical drawback of using colour features is that they are sensitive to illumination and contrast variations. This is an important issue for natural images but not for hand drawings if they are drawn and digitized under the same conditions.

For all those reasons we based our method on colour information only. Given a comic page, we first ask the user about the position of one object example and then we perform an exhaustive search in all the pages of the same album or collection.

The proposed method can be summarized as follows (for a given comic book):

- Compute and apply a reduced colour palette to all pages
- Get the object colours from input query
- Compute the query descriptor
- Retrieve similar objects

The retrieval (search) step can be extended to other albums from the same collection (similar comic characters).

4.5.1 Colour quantization

Once the user has highlighted the query region, we reduce the number of colours of all the images of the album (including the query region) according to a colour palette P_c. Colour reduction or quantization generally involves two steps. The first step consists in choosing a proper colour palette and the second step is to reconstruct an image by replacing original colours with the most similar colour from the palette.

Before quantizing the image colours, we smooth the image in order to reduce the impact of printed dithering effects that sometimes appear during the printing and

scanning process, in order to obtain flat colours. We use a state of the art method for image smoothing that has shown very good results for cartoonification and image compression denoising [201]. An alternative can be used for halftoned colour print scanned at high resolution and dithering [88] (Figure 4.10).

We use K-means clustering method for colour quantization into N colours in order to compute a palette of N colours representative of the image content. The number of colour/cluster can vary according to the type of image (e.g. flat, realistic, watercolour), we fixed $N = 256$ because comics are generally colourful drawings that contain a limited number of colours, $N = 64$ or $N = 32$ can also be used in most of the cases (Figure 4.10). Then we create a quantized image where each pixel colour corresponds to the colour of the closest cluster centre. The selection of the closest centre is performed by comparing the original colour of the pixel with the three channels (RGB) of each cluster centre. This process should be applied to the full album in order to ensure that the colours will be identical from one page to another. This limited amount of colours allows describing and retrieving the object with only few representative colours.

Figure 4.10: Colour image smoothing and quantization effects for comic character extraction. From left to right columns: original image, smoothing [201] and k-means clustering ($N = 32$). Image credits: [37, 77, 94, 107].

4.5.2 Input query

A minimal user interaction is necessary to tell the system what to retrieve in the album or collection. We only ask the user to give information about the object's colours by selecting a few pixels from the query object. We propose to use a pointing device (e.g. mouse, finger on touchscreen) to click, drag and release over the object colours and make the selection of pixels (Figure 4.11). This selection can be assisted using a scribble-based tool [202].

Figure 4.11: Examples of comic character colour selection using a pointing device (click, drag and release). Selected pixels are highlighted in red, they constitute the query given by the user. Image credits: [37, 94, 124, 147].

4.5.3 Query descriptor

The query descriptor we use is a non-ordered colour vector that contains the information of presence of certain colours in the query ($D \subset P_c$). The fact of not including any spatial or quantitative information (e.g. region positions, number of pixels) allows the method to be scale and deformation invariant.

The colour subset D is composed by the N colours most represented in the query region and that are the most discriminative in the image or collection. We define two metrics for each colour of the query region in order to determinate which colour will be used for the query description. The first metric L_{1c} corresponds to the number of pixels of a colour c in the query region (colour histogram Figure 4.12) and the second one L_{2c}, the discriminability level of the colour c in the whole image.

The discriminability level aims at informing about how a given colour is relevant for describing the comic character. The optimal situation is when a colour appears a lot in the region of the character and only in this region, this would facilitate the

character spotting process. It can be formalized as a TF-IDF [25] (Term Frequency - Inverse Document Frequency) numerical statistics that allows us to identify the importance weights of the different colour values in the query image. In our application, the *term frequency* (tf) corresponds to the ratio of pixels of a particular hue value h_i in the set of hue values that compose the query region $qH = \{h_0, h_1, ..., h_n\}$. The *inverse document frequency* (idf) corresponds, for a given hue value h_i, to the logarithm of the total number of pixels in the image $|p|$ over the number of pixels of hue value h_i in the image $|p| \cup |h_i|$. The score for each hue value h_i in qH corresponds to $L_{2c_i} = tf_i * idf_i$ and the N hue values that have the N highest scores are used as colour descriptor of the query region $D = \{h_0, h_1, ..., h_N\}$. See Equations 4.7 and 4.8.

$$tf = \frac{nbPixelMatchInQuery}{nbPixInQuery} \tag{4.7}$$

$$idf = \log \frac{nbPixInImage}{nbPixelMatchInImage} \tag{4.8}$$

The discriminability level L_{2c} allows us to ignore colours that are dominant both in the query region and in the rest of the image. For instance, black and white colours are often in the query region but also in many contour and background regions which will not help the retrieval process if we include them in the query description (they are not discriminant colours).

Each value of L_{1c} and L_{2c} is normalized as a percentage between 0 and 1, the score for each colour is the average of both terms: $Score(c) = 1/2 * (L_{1c} + L_{2c})$ (Figure 4.12).

4.5.4 Object retrieval

Comics have a particular structure that allows different approaches for object retrieval. An interesting point is that same instances of objects often appear in different panels and therefore several times in each page of the album.

First, for a given page, we consider each colour value in the descriptor and compute the corresponding colour mask (Figure 4.13). Second we find the object position in each page by using a sliding window approach at different window sizes and for each mask. The window sizes are defined according to the user query size (maximum between the query width or height) which gives an information about the definition of the document $S = \{s_0, s_1, ..., s_n\}$ in order to retrieve the object at different scales. Each window position is computed with 50% per cent overlapping with its four neighbouring windows.

We define the detection confidence C_{w_i} according to the number of identical colours between the query descriptor D and each sliding window position $W = \{w_0, w_1, ..., w_n\}$ (Equation. 4.9). This is equal to the cardinality of the intersection between the two colour sets $D \cap w_i$. The confidence is maximal when all the colours

Figure 4.12: The left part represents a query example after the colour reduction process and its original image. Top-right part shows the colour histogram of the 6 first colours with the highest number of pixels in the query as a percentage from the total number of pixels in the query (L_1). Middle-right figure represents the corresponding discriminability level for each colour according the other pixel in the page. Bottom-right figure represents the corresponding query descriptor of $N = 3$ colours. Image credits: [94].

in the query descriptor are contained by the window (high probability of the object to be at this location). The confidence C_{w_i} is defined as follow:

$$C_{w_i} = \frac{|D|}{|D \cap w_i|} \tag{4.9}$$

where C_{w_i} is the confidence value for the region w_i, $|D|$ the cardinality of the colour set D and $|D \cap w_i|$ the cardinality (number of elements) of the intersection between D and w_i .

The confidence value is compared to a threshold value T to decide whether the detection is correct or not. See the detection part of Algorithm 2. The threshold T is discussed in Section 6.9. By using a multi-scale approach, the same region may be detected at different scales. In order to keep only the smallest region that includes a minimal amount of background information, we compute a score p for each detected regions, which is the percentage of pixel that are of a colour from the descriptor. A post processing step removes multiple object detections by keeping only the regions (windows) that are not overlapped by other regions with a higher score p (best detected). See filtering part of Algorithm 2 and the illustration Figure 4.14.

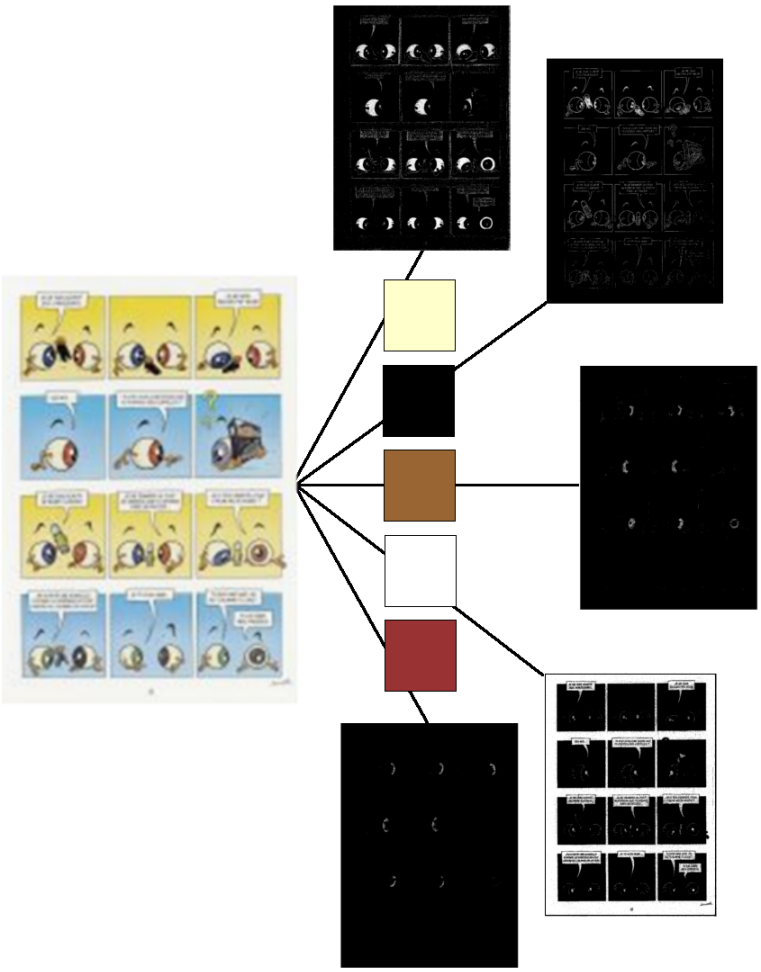


Figure 4.13: Example of colour mask corresponding to the 5 first colours in the top histogram of Figure 4.12. Image credits: [94].

4.6 Conclusions

In this chapter we have proposed independent extraction methods for panel, text, balloon and comic characters. The advantage of such independent processing compared to the sequential approach presented in the previous chapter is the non-propagation of errors throughout the process and the flexibility for adapting each contribution to different configurations or other domains.

The panel extraction method is simple, fast and efficient for comics layout that have disconnected panels as the method relies on outermost contour analysis. This method is evaluated in Section 6.3. The proposed approach for text extraction consists

Algorithm 2 Object retrieval

```
//Detection
for each s in S do
  for each w in W do
    if C(w) >= T then
      add w to the detected region list R
    end if
  end for
end for

//Filtering
for each r in R do
  if r does not overlap better detected region then
    r is a candidate region
  end if
end for
```

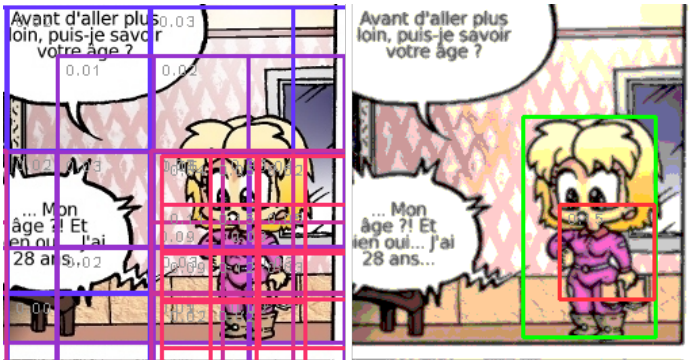


Figure 4.14: Multi-scale spotting example. The left image shows the detections at different scales, the colouration depends on to the p value, blue to red for low to high. The right side image shows the remaining best detection after the filtering step. Image credits: [37].

in an adapted bi-level segmentation followed by a filtering process of three steps based on contrast, topology and local similarities to classify the connected-components as graphics or text region. A final text line level verification validates the textual content of the candidate regions using an OCR system. The method does not require previous knowledge about the image content unlike the sequential approach proposed Section 3.2 that is implicitly linked to the panel regions. This method is evaluated and discussed Section 6.4.

Independent balloon extraction and classification methods have been presented.

The balloon extractor relies on topological and spatial relations similarly to the text extraction method. In fact, balloon and text extractions are often correlated in the literature because those regions are often associated to convey most of the textual information of comics. As mentioned earlier, tail detection without previous balloon segmentation makes no sense because it is a specific region of the balloon contour. This method is evaluated in Section 6.6. Unsupervised comic characters extraction is a real challenge, here we proposed a user-defined query-based approach but it can be replaced by a comic character region of interest, automatically computed from surrounding information (context) (Section 3.5). In each extraction process, we strive to produce indicators of confidence, conscious that it could be useful for a higher level processing system such as that presented in the next chapter.

In the next chapter, we propose to go one step forward by using a system able to measure and improve the coherence of the information being extracted. This high level approach is also used for understanding the image content and to retrieve the semantic relations between elements (context). We propose two models to embed the domain knowledge of both comic books and image processing. These models allow consistency analysis of extracted information and inference of the relations between all the extracted elements such as the reading order, the type of text (e.g. spoken, onomatopoeic, illustrative) and the relations between the speech balloons and the speaking characters.

Chapter 5

Knowledge-driven analysis

In this chapter we address the problem of understanding comics by automatically improving the consistency of the extracted elements according to the domain knowledge and by inferring new information and launching further processing iteratively. Detecting these elements and establishing their relationship is crucial towards complete comics understanding. Using this approach we propose a contribution towards the adaptability of an interpretation system, which corresponds to a hard belt of the literature. Low level and high level information is used in a complementary and unsupervised way, to progressively identify the different elements of a digitized comic book image and their semantic relations.

5.1 Introduction

Document image analysis is an active field of research which can attain a complete understanding of the semantics of a given document. One example of the document understanding process is enabling a computer to understand a comic strip story. In this chapter we propose a knowledge-driven system that can interact with bottom-up and top-down information to progressively understand the content of a document at a semantic level (Figure 5.1).

The proposed framework is based on interactions between low and high level processing in order to reduce the challenging semantic gap. The knowledge about low and high level processing is stored in ontologies that are fed by the low level image content extractors. The knowledge base is associated with properties and constraints that form an ontology. An expert system uses the ontology to assert the relations between regions and an inference engine to deduct new information in order to perform further image processing. The main difficulty is to extract but also to model the diversity of styles, format, definition and the differences between design and printing techniques (Figure 2.4).

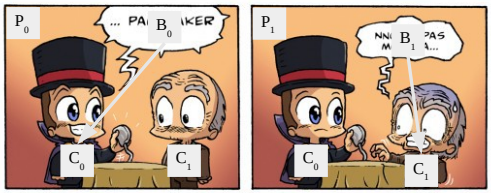


Figure 5.1: Example of semantic information understanding. The left panel P_0 represents a comics character labelled as C_0 saying the content of balloon B_0 to another character labelled as C_1. In the right panel P_1, the character C_1 is saying B_1 to C_0. Image credits: [39].

This work is the result of an extensive collaboration with Clément Guérin, a Ph.D. student working on knowledge representation applied to comic book contents. We jointly designed the knowledge model and the interactions between the low and high level processing. Further details about the knowledge models, inference engine and the interactions between the elements can be found in his thesis [62].

5.2 Proposed models

In this section, we introduce the ontologies developed for automatic analysis and understanding of comic book images. According to the state of the art review presented in Section 2.3, only one previous work proposed an ontology of comics from a philosophical approach [122]. Guérin *et al.* already published some work concerning the inference of the reading order, for panel in the page and balloon in the panel [61,62]. Here we focus on semantic relations within the panel region which is a more recent work in the team. First, we present a first ontology formalising the concepts related to image processing, data type, the inputs and outputs. Secondly, we conceptualise the field of comic books oriented to digital image analysis. Finally, we explain how these ontologies communicate with each other, so they can be used together.

5.2.1 Image processing domain

We initially developed a model formalizing the primitive notions of image processing. This model does not deliberately include any generic concept for comics in order to be used in other areas related to document analysis systems. It provides support for structuring data from image processing algorithms for subsequent semantic enrichment. The data produced by such algorithms are generally limited to spatial data, lines, areas, which we will reference under the unique term of *regions of interest* or *ROI* (Region Of Interest) later in this document. They are defined by their Cartesian coordinates in the orthonormal image from which they are extracted. From this coarse analysis, we define the first two concepts of our model: *Image* and *ROI*. Then, a third concept called *Extractor* is created to represent the low level algorithms that

are getting the ROI from the image pixels. Figure 5.2 shows a visual representation of these concepts altogether.

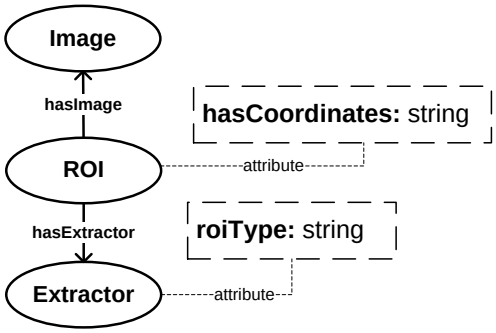


Figure 5.2: A representation of the image model involved in the expert system. Concepts are represented by the oval-shaped items; the arrows are the object properties linking them to each other. The dashed rectangles contain the data properties of the concepts they are attached to.

Image concept The concept of *Image* models the notion of image as a digital object, the input material for image processing systems. This is the most general concept from which are derived the other concepts of our ontology. An image processing algorithm works by directly manipulating the pixels of the image, its elementary components. It can analyse the pixels independently, or grouped in sets of pixels but, in the end, the value and the intensity of each pixel have their importance in the analysis process, and can affect the output results. Note that for two visually similar images, the individual value of each pixel can vary greatly from one image to another (e.g. high definition, quantization, dithering). Furthermore, the encoding of the visual information in each pixel also varies depending on the image format and the degree of compression applied. It is important to integrate this information into the model as they define the context of the processing tools. They greatly help the understanding of the result quality.

Region Of Interest concept The concept *ROI* derives from the concept *Image* (a region of interest is part of an image). This concept formalises the notion of region of interest as perceived by the image analysis. It is composed of connected pixels with visual features corresponding to a certain class of desired visual objects. The fact that a region of interest is a two-dimensional surface, allows us to represent ROIs as polygons.

Region extractor concept A region of interest identified by an algorithm of image segmentation corresponds to the proposal by the latter of a spatial area of the image containing one of the elements that the extractor is designed to detect. This region can actually contain an element of this type or contain anything else (false

alarm). The accuracy of the segmentation algorithm is dependent on the characteristics of the image, the complexity of the region to extract, the way the parameters are adaptively tuned and of course the quality of its implementation. Regions of interest are produced only as proposals requesting to be checked.

We call such image analysis systems *Extractors* later in this document. The notion of extractor is incorporated into our model as *Extractor*. Linked to the *ROI* concept, it expresses a close connection between the image processing algorithm and the regions of interest that are produced from an image with given characteristics (Figure 5.2). This set of concepts provides an initial framework for the integration of a ROI in a more complete system where it will have to interact with other entities. A region of interest from an image is produced by a single extractor.

5.2.2 Comics domain

We present in this section a conceptualization of the comics domain and its ontological formalisation. This conceptualisation has been designed by keeping in mind the purpose of use in the context of image analysis. See [62] for its adaptation to other applications.

The comics domain contains two different information levels. First, the bibliographic information, which is common with any book, about the collection, artist name, International Book Number (ISBN), language, number of page and so on. Second, the information about all the elements that compose a comic book page (e.g. panel, balloon, text and character), their relations and the corresponding semantic.

Album and pages As seen Section 1.1, a comic book page is defined as a series of images conveying a message. These images are spatially juxtaposed in a plane that is called *page* or *board*. In classical comic strips, a story is told through an ordered succession of pages. These are represented by the printed pages grouped in an album, which can be from a collection telling a larger story. The webcomics also make use of pages, materialised by digital images. In the case of printed comics, a scanned image may extend over one or two pages, rarely more. A webcomic page is always represented by a single image.

The first two concepts are introduced as *Comic* and *Page*. These two concepts are related by the property *hasPage*. A page is only part of one album.

Bibliographic information of each album are represented by attributes associated to the concept *Comic*. The album title, the collection (if it exists), its authors, date of publication and ISBN can be given via the corresponding attributes. The reading direction (left to right or right to left) may be indicated through a Boolean attribute called *right2Left*.

Page content We consider that a page consists of panels, which contain drawings such as comic characters and speech balloons. We initially chose to focus on the

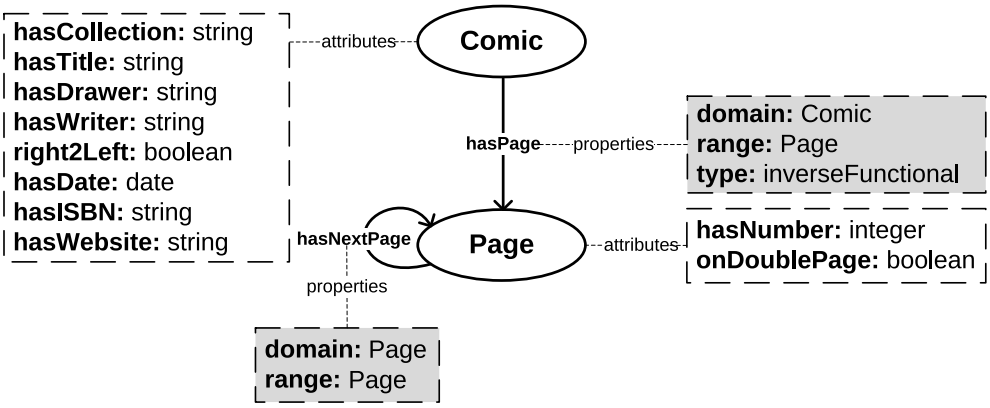


Figure 5.3: Concepts *Comic* and *Page* with their corresponding relations and attributes.

relationship between these two types of content. The consideration of other elements is beyond the scope of this work, although, extracting relevant terms from WordNet might be interesting as well [209]. Balloons contain text lines, embodying the words of the comic characters and the narration of the story.

Panels are represented by the concept *Panel* with the attribute *hasRank* which indicates the reading order in the corresponding page.

Balloons or phylacteries, whether they are spoken, thought or narrated, are represented by the concept *Balloon*. In a similar manner to panels, the balloons must be read in a appropriate order defined by their spatial position in their corresponding panel. We consider *attached* and not *overlapped* panels because, according to the comic book styles, the balloons are not necessarily spatially positioned within a panel. Sometimes they are slightly outside or straddle several panels. Instead, each panel illustrates one action taking place at a fixed time. Balloons are an integral part of the staging and are attached to a given panel, independently from their relative position to this panel. Their position in the reading order of a panel is defined by the attribute *hasRank* and the property *hasNextBalloon* binds balloons together.

The balloon tail is represented as the concept *Tail*, and its direction by the attribute *hasDirection*. A balloon may be related to several tails through the property *hasTail*.

Text lines a designed by the concept *TextLine*. They are grouped inside balloons with a reading order from top to bottom. Similarly to panels and balloons concepts, the attribute *hasRank* indicates their position in the reading order inside the balloon through the property *hasNextTextLine*. Text transcription is stored via the attribute *hasText*.

The concepts *Panel*, *Balloon*, *Tail*, *TextLine* and *Character* are disjoint, each element can only be an instance of one of them. Figure 5.4 illustrates the addition of

these concepts to the ontology.

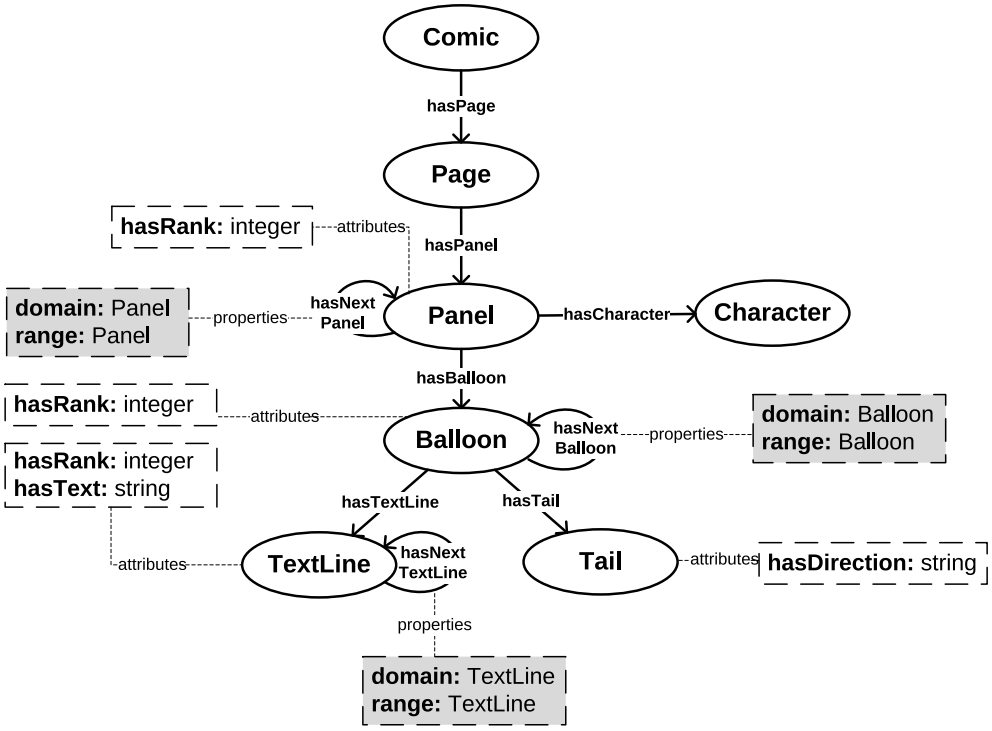


Figure 5.4: Integration of concepts *Panel*, *Balloon*, *Tail*, *TextLine* and *Character* to the initial model Figure 5.3.

Several properties are introduced into our ontology to represent the links between the various components of a panel. A panel being relative to the page, the property *hasPanel* binds an instance of *Page* to an instance of *Panel*.

Properties *hasBalloon* and *hasCharacter* are formally defined and represent the existing membership between, on one hand, a box and, on the other hand, an instance of *Balloon* and *Character*. The property *hasTextLine* represents the link between a text line and a balloon.

Specialisation of the content The semantic level of the presented concepts remains at a degree of granularity quite crude, here we present how to refine it. Balloons might be categorized into two subsets according to their relation to comic characters or not. On one hand the balloons emitted by characters (spoken or thoughts) or elements of the scene (radio, television, etc.), and on the other hand, narrative balloons. The shape of the speech balloon varies from one author to another. One feature which seems to be a consensus to discriminate narrative balloons from others, is the presence or not of a tail pointing to the source of the sound. The concepts *SpeechBalloon*

and *NarrativeBalloon* are introduced to represent balloons equipped with a tail or not respectively.

The semantic of text lines in these newly specialised balloons can then be refined accordingly. Some text lines are carrying elements of speech, while others are for storytelling. The two corresponding concepts are *SpokenTextLine* and *NarrativeTextLine* respectively. They are simply defined as text lines belonging to an instance of *SpeechBalloon* or *NarrativeBalloon*.

Speech balloons are usually issued by a character present in the panel. The link between a character and a speech balloon is expressed through the property *says*, whose domains *Character* and range *SpeechBalloon*. The concept *Speaker* represents a character which is emitting a speech balloon.

Figure 5.5 illustrates the relations introduced for the concepts *Balloon*, *TextLine* and *Character*.

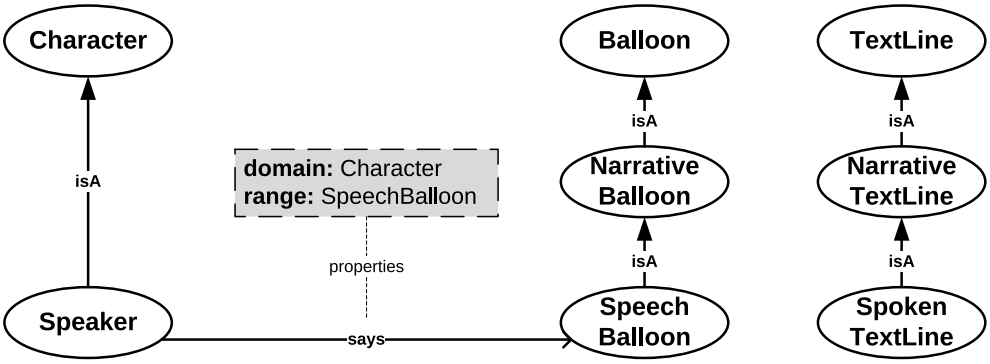


Figure 5.5: Specification of concepts *Character*, *Balloon* and *TextLine*

5.2.3 Model interactions

In this section we present the interactions between the image and comics ontologies so that they can communicate and combine their reasoning capabilities. We call $\mathcal{O}_{image}$ and $\mathcal{O}_{comics}$ the ontologies of image and comics respectively presented in Sections 5.2.1 and 5.2.2. Figure 5.6 illustrates the interactions between the ontologies.

The image and comics models are linked through two bridges. First, the *Image* concept from $\mathcal{O}_{image}$ and the *Page* concept from $\mathcal{O}_{comics}$ are made equivalent. This ensures that all extracted content related to an image is equally related to a corresponding page in the comics domain $Page \equiv Image$.

Second, the classes $Cl = \{Panel, Balloon, Tail, TextLine, Character\}$ are defined as equivalent to the corresponding set of regions of interest $roiType = \{panels, balloons, tail, textlines, characters\}$ (Equation 5.1).

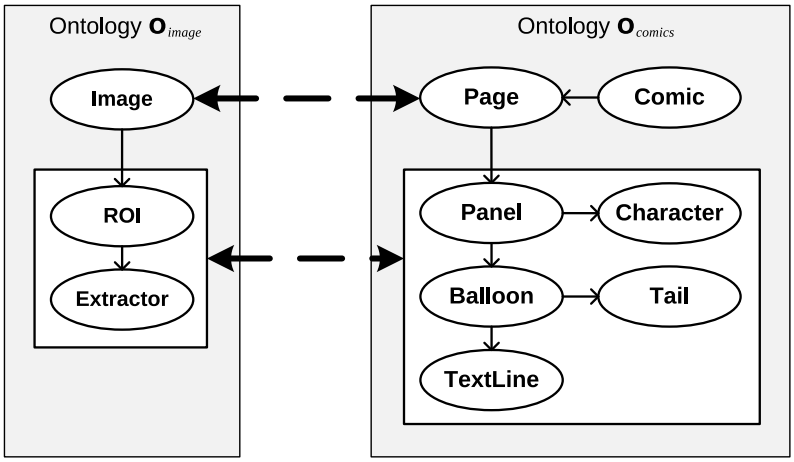


Figure 5.6: Interaction between the two ontologies.

$$Cl_i \equiv \text{ROI } \textbf{and } (hasExtractor \textbf{ some } (roiType \textbf{ value } Sr_i)) \tag{5.1}$$

5.3 Expert system for contextual analysis

In this section we refer to the algorithmic part related to image processing as *low level* processing. The proposed system composed by the developed ontologies and an inference engine (*high level* processing) is called *expert system* for clarity. A dynamic communication is maintained between low and high level processing of the overall analysis system, this section will detail its operation.

5.3.1 Interactions between low and high level processing

We consider the ontologies $\mathcal{O}_{image}$ and $\mathcal{O}_{comics}$ developed as integral parts of an expert system that provides an interpretation framework for low level image content. The purpose of the expert system is to interact with the low level iteratively in order to progressively understand the content of an image, moving from simple to more complex elements. This approach is similar to [159] except that in our case the definition of the complex object is not a composition of simple objects but context-driven.

The expert system, represented by the diagram in Figure 5.7, includes both ontologies forming our knowledge base. This knowledge base, once populated by data from the low-level system, is queried by an inference engine (e.g. Racer [64]] or Pel-

let [166]) to extract logical conclusions, depending on the data and their consistency compared to the formalised knowledge.

These logical conclusions may include a validation of extracted elements, rejection or creation of elements referred to the image processing system.

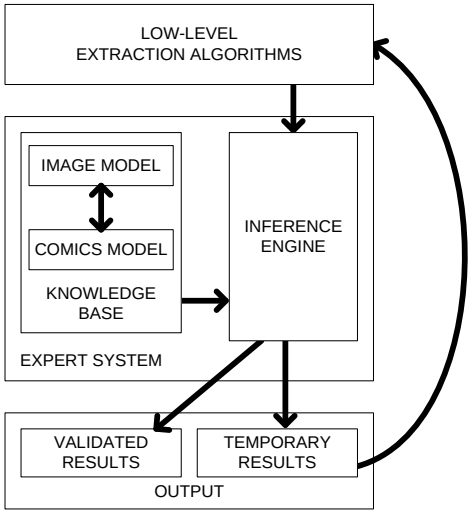


Figure 5.7: Generic representation of the expert system and the relationship between knowledge base, the inference engine and the low-level algorithms.

The low level algorithms have been designed to extract specific information from the whole image or a specific region. Low and high level systems interact in a loop to feed the knowledge base until there is a complete and consistent understanding of the document, according to the knowledge domain. Figure 5.8 illustrates this loop interacting between the two levels of information.

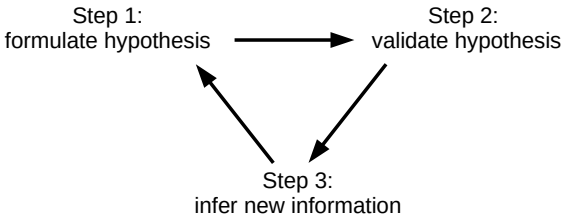


Figure 5.8: Process loop of the framework.

In step 1, the low level system propose to the expert system assumptions about a first set of segmented regions in the image. These regions are labelled according to their supposed type according to the extractor (panel, balloon, text, etc.). In the second step, the expert system evaluates these assumptions, validates correct ones and deletes others. Note the deletion is the simplest action we can perform on

invalid information, other scenarios such as switching or changing element labels have to be considered as well. In the third step, new information is inferred from valid assumptions, put in perspective of the domain knowledge. These are referred to the corresponding low level extractor that will use them, during the following iteration, to extract more complex elements such as the comic characters. This loop can be run as many times as new information is discovered.

5.3.2 Constraints for the low level extractions

In our model, knowledge is modelled though the categorization of each element composing a page, combined with a set of topological relations with these elements. In our context, the elements that compose a given image I are panels P, balloons B, tails Q, text T and characters C, as well as the set of topological and semantic relations between them. Because comics, as an art form, do not follow any strict specifications it is really hard to build a perfect model which is valid for all kinds of comics. There are some instances of comic books without balloons or without panels. If webcomics are also considered, then a comic is not even necessarily composed of pages. A model that would be true for every type of comic book would be too general to be of any use in this work. Instead we define a general comic book model with more constrained properties that represent a large subset of comics (Franco-Belgian, Japanese and American). The main advantage is that it can be adapted to any kind of document images by defining properties according to the application domain. We define the general properties of comics as follows:

- A panel P is related to one and only one comics page
- A balloon B is related to one and only one panel
- A character C is related to one and only one panel
- A text line T is related to one and only one balloon B

Despite the fact that authors are free in their layout choices, they follow general use conventions widely adopted by comic book's authors in order to avoid confusing the reader [46, 91]. The depicted elements and their placement in the page must be clearly identifiable at first sight, meaning, for instance, that text is contained by the balloons which are included inside panels just like the characters. Whereas one can find some instances of balloons breaking out of their panel, these are usually kept to a minimum.

Therefore, the term "related" refers to the situation where an object is overlapped (a fortiori, contained) by another over a *significant* proportion of its surface. In the case of multiple intersections, only the smallest container is considered. When the element is fully contained in several other items, the smallest container is consequently the direct container (e.g. a text line must be considered as being included inside a balloon before being included inside the panel containing that balloon).

5.4 Processing sequence

The expert system asserts the extraction of simple elements such as panels, text, balloons and tails in order to infer speech balloons before searching for more complex elements (e.g. comic book characters) based on the context defined by the simple elements and their relations. This can be demonstrated with the first two iterations of the process loop (Figure 5.8).

The first iteration treats simple elements such as panels, balloons and text lines. They are called simple because they have a relatively regular structure. During the second iteration of the process loop, the detection of the comic characters, more complex in their visual diversity, is treated.

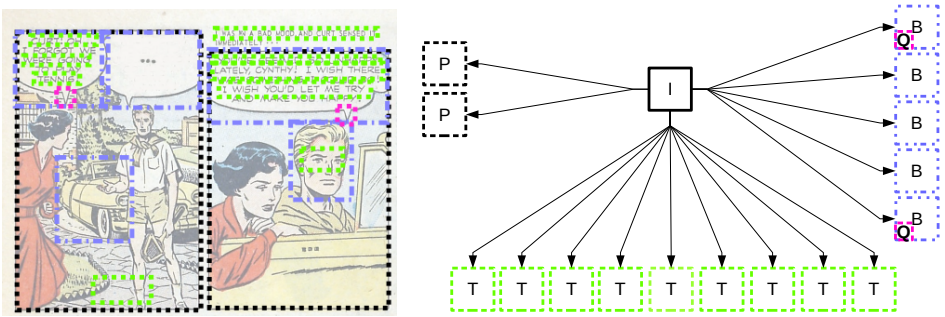
The different stages of the process are illustrated through the couple of panels presented in Figure 5.9.

Figure 5.9: Original panels used to illustrate the different stages of the processing sequence. Image credits: [142].

5.4.1 Simple element extraction

Iteration 1 - step 1 (hypothesis) The initial extraction of panels, text and balloons feeds the knowledge base. All the elements are extracted independently using the method proposed in Chapter 4. All these elements are assumptions to be validated by the expert system. In Figure 5.10, dashed elements represent the initial hypotheses and each colour a result from a different extractor. Note that extraction errors can take place at this stage and that these errors can be recovered by the system at a later stage.

Iteration 1 - step 2 (validation) At the second step, the hypotheses proposed by the low level system are compared to the constraints formalized in the ontology. The extracted regions are first categorized as panels, balloons and text lines thanks



(a) (b)

Figure 5.10: Initial hypothesis (dashed elements) about the content of a given image I after the initial extractions of panels P, text T and balloons B with tails Q.

to the rules presented in Section 5.2.3, interpreted by the inference engine. Then each element e is linked to its direct container E. This is selected from the set of extracted regions as described in Section 5.3.2. If the system is unable to find a container for an element e from the set of extracted regions, the page is then considered as the container of e.

In order to validate the assumption of all the elements in E, they are simply considered as instances of the concept ROI_i. We voluntarily do not take into account of their class in the ontology $\mathcal{O}_{bd}$. This has for effect to introduce inconsistencies between the type of container and the domain of a property resulting from *hasContent*, making the model inconsistent. These inconsistencies allow highlighting possible errors made during the low-level processing. The inference engine is started again on the ontology and inconsistencies are treated one after the other.

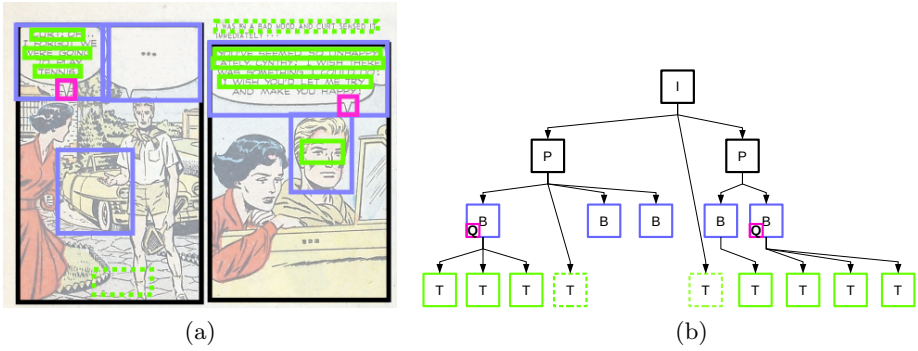
In this work, we focused on optimizing the reliability of the results. We have chosen to focus on increasing the extraction precision rather than completeness by deleting the elements that did not fit the proposed model. The detection of misclassified elements (e.g. panel actually being a balloon) and the proposition of missed elements (e.g. a missed balloon around a group of text lines) are both short term perspectives of this work. For the time being, our system can handle, without being limited to, the following inconsistencies:

- **A page (p) contains a balloon (b) or a text line (t)**: b or t is deleted.
- **A panel (p1) contains a panel (p2) or a text line (t)**: p2 or t is deleted.
- **A balloon (b) contains a panel (p)**: if p contains some balloons and b does not contain any text lines, b is deleted, otherwise, p is deleted.
- **A balloon (b1) contains a balloon (b2)**: if b1 does not contain any line

then b1 is deleted, otherwise b2 is deleted.

- **A text line (t) contains a panel (p)**: if p does not contain any balloon then p is deleted, otherwise t is deleted.
- **A text line (t) contains a balloon (b)**: if b does not contain any text line then b is deleted, otherwise t is deleted.
- **A text line (t1) contains a text line (t2)**: if t1 contains other lines then t1 is deleted, otherwise t2 is deleted.

The term "related" refers to the relation of membership of an element to another, as described in Section 5.3.2. Figure 5.11 illustrates the current organisation of the extracted elements. They have been structured according to the model definition Section 5.2.2. Validated elements are illustrated in solid lines while others still dashed.



(a) (b)

Figure 5.11: Validation of the hypothesis using the constraints of the knowledge base. Valid elements have a solid border.

Iteration 1 - step 3 (inference) At this step of the first iteration, the inconsistencies have been resolved and the remaining elements are in accordance with our model. They are organized in the hierarchy of concepts in $\mathcal{O}_{comics}$. The ontology being consistent, the inference engine is able to classify instances of *Balloon* and *TextLine* into *SpeechBalloon*, *NarrativeBalloon*, *SpokenTextLine* and *NarratedTextLine* according to their properties and constraints. The classification of the region instances is shown in Figure 5.12.

5.4.2 Complex element extraction

This is the beginning of the second iteration of the process (Figure 5.8) where we focus on more complex elements such as the comic characters. At this point, the expert system has some information about the content of the page from the analysis carried out during the first iteration.

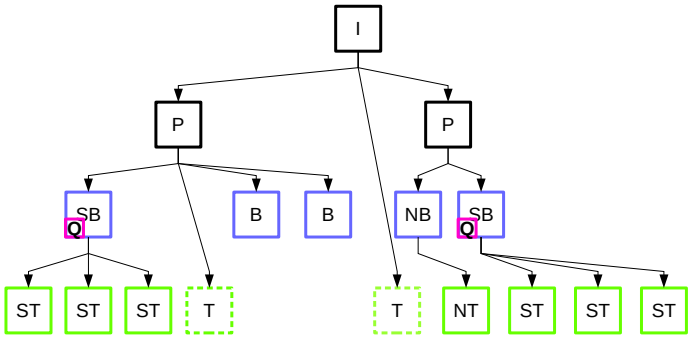


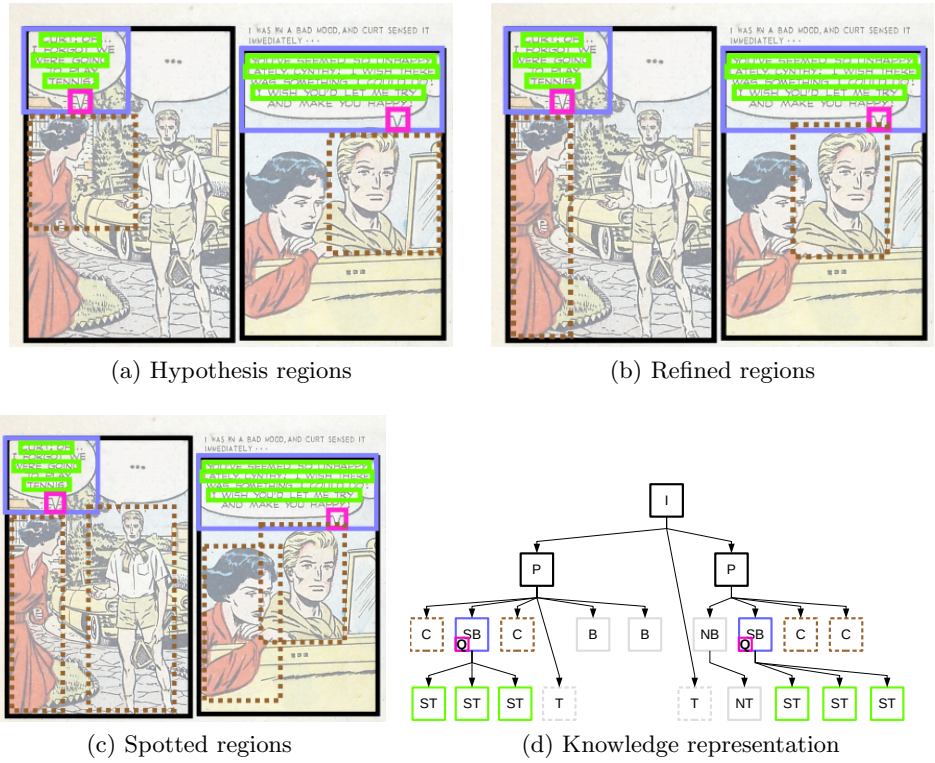
Figure 5.12: Inference of the speech balloons SB, narrative balloons NB, spoken text lines ST and narrated text lines NT using the semantic properties and constraints of $\mathcal{O}_{comics}$.

Iteration 2 - step 1 (hypothesis) The information discovered during the first iteration of the process is used to formulate hypotheses about the probable location of the characters in the panels in order to guide their detection by the low level processing. We first focus on characters that are emitting at least one speech balloon (considered as main characters). It is reasonable to consider that the speech balloon and the character emitting the balloon belong to a same panel. It is therefore possible to restrict the search space of the character inside the panel. In addition, the position of the character in the panel can be estimated according to the direction pointed by the tail which is a characteristic of speech balloons (Figure 5.13a). The estimation of the position of the comic character from the tail position and direction have been presented in Section 3.5. The estimated regions of interest are given as seeds to the image processing algorithm (extractor of characters) which refines the regions according to the image content (Figure 5.13b). The last step from the low level processing consists in spotting all the comic characters from the estimated examples (Figure 5.13c).

The method used for ROI refinement has not been completed during this thesis, a good research line could be to fit the estimated ROI to the set of regions that are mainly overlapped by the ROI and exclude the others. The spotting approach is relevant for retrieving non speaking characters as demonstrated Section 4.5.

Iteration 2 - step 2 (validation) The validation of the new regions is performed in the same way as for the first iteration. This new batch of extraction is submitted to the constraints of the ontology as following:

- **A page (p) contains a character (c)**: p is deleted.
- **A balloon (b) contains a character (c)**: c is deleted.
- **A text line (t) contains a character (c)**: c is deleted.



(a) Hypothesis regions (b) Refined regions

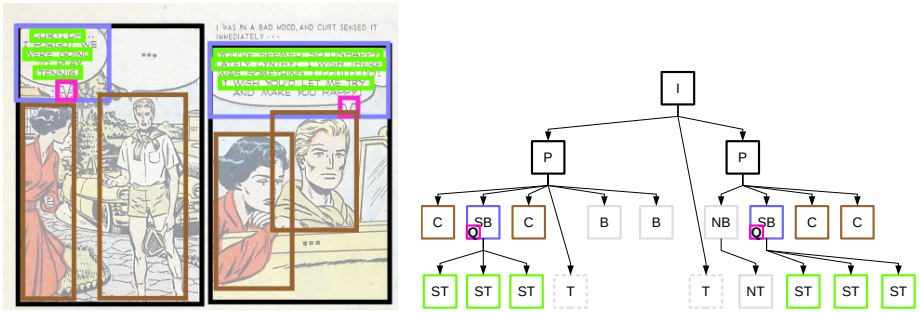
(c) Spotted regions (d) Knowledge representation

Figure 5.13: Hypothesis, refinement and spotting of comic characters regions C from the speech balloon SB regions. The regions that are not related to ST have been shaded in the graph and removed from the image to make it more comprehensible.

- **A character (c) contains a panel (p), a balloon (b) or a text line (t)**: c is deleted.
- **A character (c1) contains a character (c2)**: c2 is deleted.

Validated elements are represented with a solid line Figure 5.14.

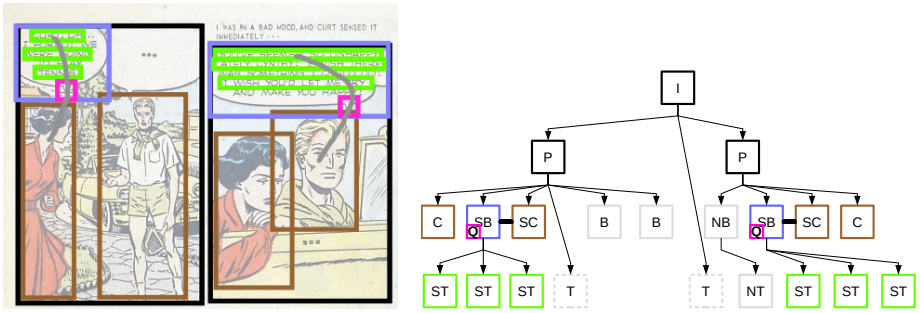
Iteration 2 - step 3 (inference) The final step of this process is to deduce from all validated characters those that are actually linked to speech balloons. Among the validated characters, we consider as being potential speakers those who intersect with the initial estimation of region of interest (Figure 5.13a). An virtual straight line is drawn from each speech balloon tail tip in the direction indicated by the tail to its related panel border. The first region of a potential speaker that it touches is considered to be the source of the speech balloon (the emitter). This relation was asserted into the ontology with the property *isSaidBy*, between the selected character



(a) (b)

Figure 5.14: Validation of the character regions C by the expert system and the corresponding image I.

and the corresponding balloon. Since the range of this property was set to the concept of *speaker*, it automatically classifies the character instance involved into this class (Figure 5.15).



(a) (b)

Figure 5.15: Inference of the two speaking characters SC and their corresponding semantic links to speech balloons SB regions. The two semantic links are represented by a grey stroke in the image over the regions concerned and a non-oriented horizontal edge in the graph.

At the end of the two iterations we obtained both a topological and a semantic description of the image content, illustrated in a single graph here. Further iteration could be processed by extracting other low level elements such as faces or vehicles and by adding extra domain knowledge.

5.5 Conclusions

We presented a new framework for understanding documents that can interact with low and high level information suitable for semi structured and complex background

documents such as comics. Several key improvements to information extraction and processing methods have been developed.

Two ontologies have been presented in this chapter, the first one is formalizing the concepts implemented for image analysis and the second one for formalizing the comics domain. The image processing ontology has the advantage of being completely independent of any application. The proposed comics ontology is composed of concepts that reflect the classic structure of a comic book. This conceptualization has been developed with the target of image analysis interfacing the image ontology through equivalence relations between some of their concepts.

The proposed system provides a novel generic and unsupervised approach for comic book document understanding mixing visual and semantic information. It relies on an inference engine that interacts with the two proposed models. We detailed a use case for comic character localisation associated with their corresponding speech balloons, taking into account the spatial organisation of the rest of the elements in an image. One limitation of the proposed system is the validation process which only suppresses elements until having a consistent representation of the knowledge. An interesting line of research could be to compute a cost, in terms of number of inconsistencies that are created or solved, of changing the class from one type to another one (e.g. changing a region given by the panel extractor as balloon and *vice versa*). This solution would permit to apply the changes that have the minimal cost, including deletion as the last option, in order to make the knowledge representation consistent while minimizing the loss of extracted information.

Another research axis would be to continue iterating the process loop with other low level extractors in order to retrieve other information such as other objects in the panels and sound effects. In addition, the expert system could also be used to improve text extraction and recognition using system feedback in order to automatically extract open speech balloons from validated text locations. A semantic tag (e.g. speech, thought, exclamation) can also be given to the speech text regions according to a deeper analysis of the type of balloon presented Section 4.4.2.

In the next chapter, we are going to put to practice the different contributions presented in this thesis and compare them to other methods from the literature. The dataset is first introduced along with the metrics we used to evaluate information retrieval.

Chapter 6

Experiments and performance evaluation

With the analysis and processing of data comes the need of the output results evaluation. Traditionally, this evaluation is done by validating the results of an algorithm against a ground truth that represents what an ideal output should be [51, 58, 167]. Ideally, such a ground truth is made publicly available so anyone can challenge his own algorithm [92]. This can be applied to any kind of results from image segmentation to classification or information retrieval. In this chapter we present the dataset and ground truth we provided to evaluate comic book related works. We describe the metrics we used to evaluate each of the contributions proposed in Chapter 3, Chapter 4 and Chapter 5 and evaluate our work compared to previous works from the literature. Note that we call these three approaches Method I (independent), Method S (sequential) and Method K (knowledge-driven) respectively in the rest of this chapter.

6.1 Dataset and ground truth

Being in need of comic books material and an associated ground truth to evaluate our work, we noticed that there is not such a dataset publicly available for scientific purposes. Therefore, we decided to gather the first publicly available comic books dataset in association with several comic books authors and publishers and to build up the corresponding ground truth according to document analysis and understanding concerns. The comic book images were selected to cover the huge diversity of comic styles with the agreement of the consenting authors and publishers that have the objective to foster innovation in this domain through academic research.

It took almost one year to define which type of comics are the most interesting for researchers, meet and convince comic book authors and publishers, get copyright authorizations for the scientific community, develop a specific annotation tool and finally to hire people to manually create the ground truth.

A selection of one hundred comic pages were annotated in one day by twenty volunteers affiliated to the L3i lab. In order to provide a common basis for evaluating research work, the ground truth was published in 2013 [63] and made available to the scientific community through a dedicated website[1]. It was enriched in 2014 by adding the location of the principal comic characters and semantic information to the already annotated elements.

The content of the dataset and the ground truth construction are briefly detailed in the two next sub sections, more details are provided in Appendix C and D respectively.

6.1.1 Dataset description

Scott McCloud defined comics as "juxtaposed pictorial and other images in deliberate sequence, intended to convey information and/or to produce an aesthetic response in the viewer" [116]. This definition is intentionally broad enough to encompass the spectrum of the majority of works produced so far. The dataset composition should reflects this heterogeneity to give everyone the opportunity to compare their algorithms on a representative dataset of the comics world. We contacted authors with different comic styles and have selected a corpus of one hundred images, representing single or double comics pages.

The images were partly processed by the French company A3DNum[2] which was commissioned to digitize 14 albums. Among all the files, scanned at a resolution of 300 dots per inch and encoded in uncompressed Portable Network Graphic (PNG) format, we used 46 pages to integrate the eBDtheque corpus. The remaining 54 images were selected from webcomics, public domain comics[3] and unpublished artwork with different styles from 72 to 300 dots per inch. We encoded all the images of the eBDtheque dataset in Joint Photographic Experts Group (JPEG) format with a lossy compression to facilitate file exchange.

Hereafter we describe the characteristics of the selected pages and their content, an overview of the images is given Appendix C.

Albums The albums have been published between 1905 and 2012. 29 pages were published before 1953 and 71 after 2000. Quality paper, colour saturation and textures related to printing technique changes can vary a lot from one image to another. The artworks are mainly from France (81%), United States (13%) and Japan (6%). Their styles vary from classical Franco-Belgium "bandes dessinées" to Japanese manga through webcomic and American comics.

[1] http://ebdtheque.univ-lr.fr
[2] http://www.a3dnum.fr
[3] http://digitalcomicmuseum.com

Pages The pages themselves have very diverse characteristics. Among all, 72 are printed in colours and according to the authors and periods, there are a majority of the tint areas, watercolours and hand-coloured areas. Among the remaining 28, 16 have are greyscale and 12 are simply black and white. One album has two versions of each page, one in colour and the other one in black and white. We have integrated examples of each of them in order to allow performance comparison of algorithms on the same graphic style by using colour information or not. Five of the 100 images are double page, others are single page and 20% are not A4 format.

Panels The panels contained in the pages are of various shapes. Although most of them are bounded by a black line, a significant proportion has at least one part of the panel which is indistinguishable from the background of the page (frameless panel). Two pages consist only of frameless panels, the visual delimitation uses background contrast difference between the panel and image. Nine images contain overlapping panels, twelve contain only panels without border and several have panels connected by a straddle object.

Balloons The balloons also contain a great diversity. Some of them are completely surrounded by a black stroke, some partially and others not at all. They have a bright background with a rectangular, oval or non-geometric shape with "smooth", "wavy" or "spiky" contour in general. Most of them have a tail pointing towards the speaker, but some have not. There is text without any surrounding balloons on 33 images of the corpus.

Text The text is either typewritten or handwritten, mainly upper-case. The text lines contain 12 elements in average (Figure 6.1) and there are more than one hundred text lines that are composed by only one letter corresponding to punctuation or single letter words such as "I" or "A"; this is a particularity of comics.

Most pages are from French artworks, where the text is written in French. Only 13 pages contain English text and 6 images are in Japanese. Onomatopoeia appears in 18 pages.

Comic characters The comic characters or protagonists are specific to each album. They all have eyes, arms and legs but at least 50% are not humanoid, depending on the interpretation.

6.1.2 Ground truth information

We give all the details about the ground truth construction process in Appendix D. The annotation of visual information such as the location of panels, balloons, text lines and comic characters is detailed. Also, present the annotation of semantic information about the type of text, relationship between speech balloons and speaking characters;

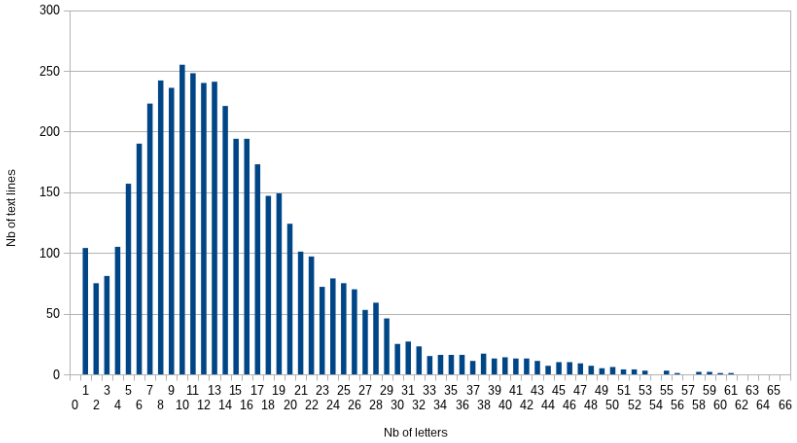


Figure 6.1: Distribution of the number of elements per text lines.

and bibliographic information about the images (e.g. page number, author, ISBN, release date). Finally, the file structure, the annotation quality assessment and the terms of use are defined.

6.2 Metrics

The contributions of this thesis are of different nature and need to be evaluated separately using appropriate metrics. Object localisation developments are evaluated using the commonly used recall and precision metrics and other contributions are evaluated using accuracy.

6.2.1 Object localisation metric

We evaluate the different extractions (panel, balloon and text regions) in terms of object bounding boxes similarly to the PASCAL VOC challenge [52]. The detections are assigned to ground truth objects and judged to be true or false positives by measuring bounding box overlap. To be considered as a correct detection, the overlap ratio a_0 between the predicted bounding box B_p and the ground truth bounding box B_{gt} (Formula 6.1) must exceed 0.5. The predicted objects are considered as true positive TP if $a_0 > 0.5$ or false positive FP (prediction errors) otherwise.

$$a_0 = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{6.1}$$

Detections returned by a method are assigned to ground truth objects satisfy-

ing the overlap criterion ranked by the confidence output (decreasing). Multiple detections of the same object in an image are considered as false detections (e.g. 5 detections of a single object counted as 1 correct and 4 wrong detections).

The number of TP, FP and false negative (missed elements) FN are used to compute the recall R, precision P and F-measure F of each method using Formulas 6.2.

$$R = \frac{TP}{TP + FN} \qquad P = \frac{TP}{TP + FP} \qquad F = 2 * \frac{P * R}{P + R} \tag{6.2}$$

6.2.2 Object segmentation metric

Object bounding box based evaluation is appropriate for surface comparison but not for detailed region extraction evaluation. In Section 4.4.2 we extract balloon contours at the level of pixel for analysis and classification purposes. In order to differentiate between error sources, we need to use a more precise evaluation framework. Here we keep using recall and precision metrics as introduced in the previous section but instead of counting the number of objects that comply with a certain criterion, we simply count the number of pixels that have been detected correctly (TP), incorrectly (FP) or missed (FN).

6.2.3 Text recognition metric

Even though text recognition would require more investigation to be fully treated, we give a first baseline evaluation in Section 6.5 using commercial OCR systems on the eBDtheque dataset. We evaluated the text detection accuracy $A_{textReco}$ at a given string edit distance between the predicted recognition and its corresponding transcription in the ground truth [63].

6.2.4 Tail detection metric

Tail tip and tail direction are not surfaces, therefore they can not be evaluated using recall and precision metrics presented Section 6.2.1. Thus, we define two accuracy metrics $A_{tailTip}$, the accuracy of the predicted position of the tail tip and $A_{tailDir}$ the accuracy of the tail direction prediction. The Euclidean distance d_0 between the predicted position of the tip and its ground truth is measured relative to balloon size (Formula 6.3). Note that we consider as incorrect the predicted positions at a distance d_0 superior to the balloon size ($A_{tailTip} < 0$).

$$A_{tailTip} = 1 - \frac{d_0}{0.5 * (B_{width} + B_{height})} \tag{6.3}$$

where B_{width} and B_{height} correspond to the balloon width and height respectively.

The direction accuracy $A_{tailDir}$ was measured according to the distance d_1 within the eight cardinal coordinate sequences defined in Section 3.1 (Formula 6.4).

$$A_{tailDir} = 1 - \frac{d_1}{8} \tag{6.4}$$

For instance if the detected direction was S (south) and the ground truth was SE (south-east) then $d_1 = 1$. Note that our method can also detect when there is no tail on the balloon contour $C_{tail} = 0\%$ (confidence equal to zero percent); in this case $A_{tailTip} = A_{tailDir} = 100\%$ if there was effectively no tail to detect or $A_{tailTip} = A_{tailDir} = 0\%$.

6.2.5 Semantic links metric

The semantic links between speech text and speech balloon are called $STSB$ and the ones between speech balloon and speaking character $SBSC$; they characterise a dialogue. They are considered true or false according to their existence or not in the ground truth. We evaluated the semantic relations $STSB$ and $SBSC$ according to the metadata in the ground truth of the eBDtheque dataset [63] called `isLineOf` and `isSaidBy`, which represent 3427 and 829 relations respectively. We defined two accuracy metrics A_{STSB} and A_{SBSC} to measure the percentage of correctly predicted semantic links (Formula 6.5).

$$A_{STSB} = \frac{nbRetrievedSTSBLinks}{nbSTSBLinks} \quad A_{SBSC} = \frac{nbRetrievedSBSCLinks}{nbSBSCLinks} \tag{6.5}$$

6.3 Panel extraction evaluation

In this section we evaluate our three approaches (Method S, I and K) and compare them to two methods from the literature. We compare our results to Arai [7] and Ho [67] that are two state of the art methods (Section 2.2.1). The first one use connected-component analysis similarly to our proposition and the second is based on growing region.

All the evaluations are performed on the 850 panels of the eBDtheque dataset (Section 6.1) at object bounding box level, using the recall and precision metrics introduced in Section 6.2.1.

Note that some of the images in the dataset were digitized with a dark background surrounding the cover of the book. We automatically remove this by cropping the image where a panel with an area greater than 90% of the page area is detected.

6.3.1 Arai's method

We have implemented as a baseline the comic panel extraction method presented in [7] except for the division line detection due to lack of detail in the original paper.

This method consists in a bi-level segmentation with an empirical threshold value of 250 followed by a connected-component extraction, binary image inversion and blob selection. The final blob selection is based on a minimal size of $Image.Width/6$ and $Image.Height/8$. From the selected blobs, a line detection approach is applied as final decision. This line detection approach is able to cut overlapped panels on a page size basis which is appropriate for pages with a single panel per strip as illustrated in Figure 6.2. The author did not share the code and we were not able to re implement this part due to lack of detail in the original paper. Anyway, the line division method works only for panels that are as large as the page which is not so common and would not have affected the result significantly (Figure 6.2).

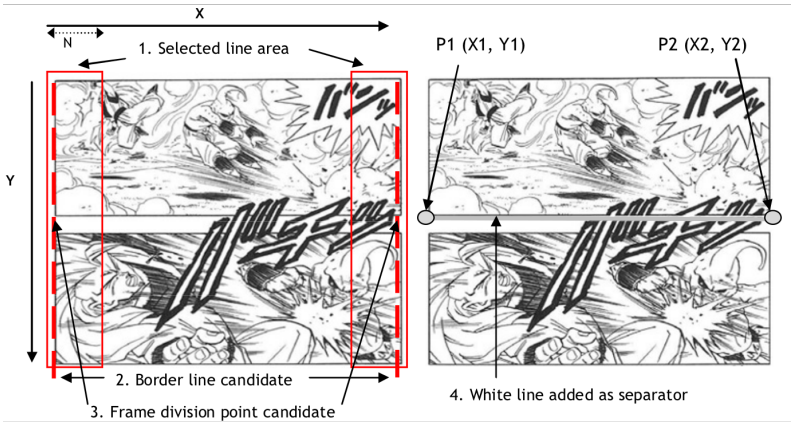


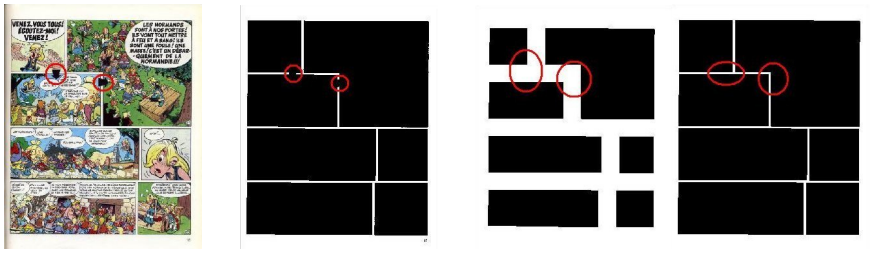
Figure 6.2: Division line detection from [7] (figure 5 in the author's paper).

This method yields an average recall of 20% and precision of 18.75% (Figure 6.4a).

6.3.2 Ho's method

We have also implemented as baseline Ho's method [67] with the original parameters and set the minimal and maximal lower brightness difference of the region growing method to 20 (not mentioned in the original paper). The image border is filled with the average five-pixel page border colour and four seeds are initialized on the four corners of the image. When the region growing algorithm stops, the background is removed which separates panel blocks. Mathematical morphology (dilation) is then applied until blocks become smaller than 1/6 of the page size and then the same number of erosion steps is applied in order to give back to the objects their initial

size (closing operation). This manipulation separates the connected panels but has the disadvantage of creating unwanted panels as well (Figure 6.3).



(a) Original image (b) Binary segmentation (c) After dilation step (d) After erosion step

Figure 6.3: Panel extraction and separation process of Ho [67].

This method yields an average recall of 49.76% and precision of 68.74% (Figure 6.4b).

6.3.3 Sequential approach

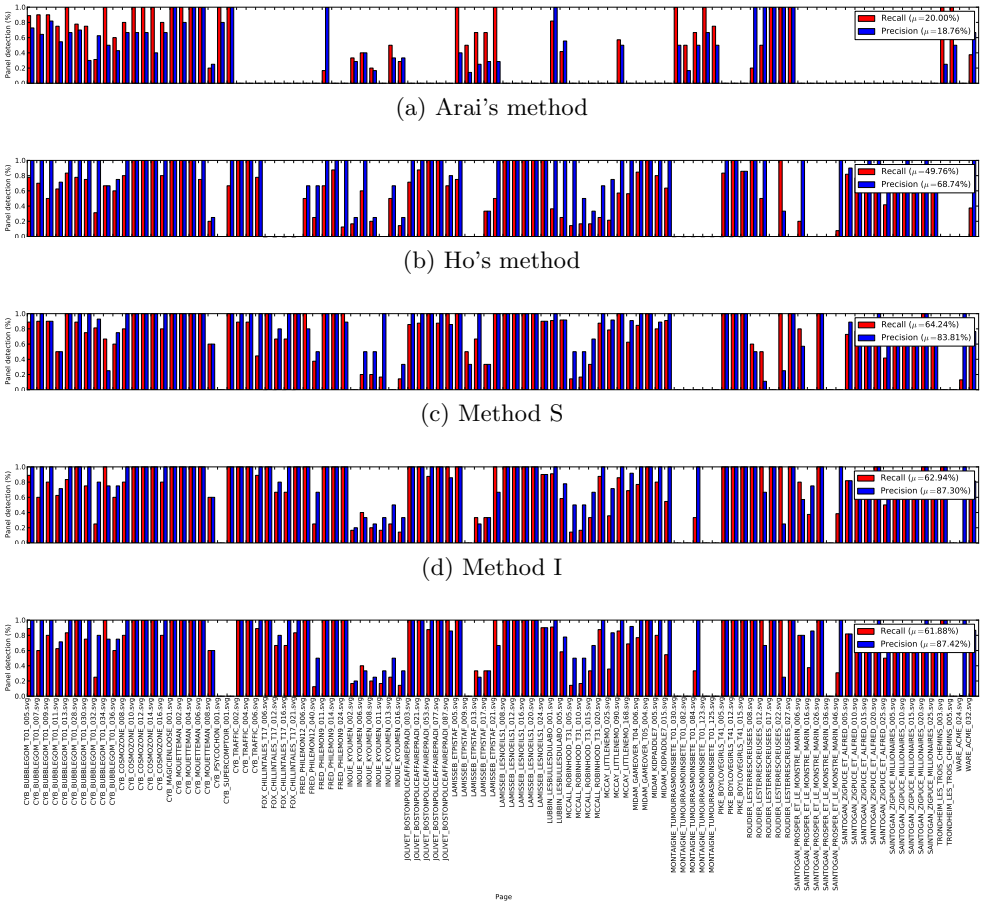
The method presented in Section 3.2 is parameter free except for the number of cluster for the k-means clustering algorithm that we fixed to $k = 3$ according to the domain knowledge of comics (Section 3.2). Note that the method also extracts text regions at the same time which do not interfere with panel (Section 6.4). This method yields an average recall of 64.24% and precision of 83.81% (Figure 6.4c).

6.3.4 Independent approach

The method presented in Section 4.2 requires only a minimal area factor. Assuming that a panel is a big region, we ignored the panel detection with a area lower than 4% ($minAreaFactor$) of the page area according to a validation on the eBDtheque dataset. This parameter avoids considering small and isolated elements (e.g. text, logo and page number) as panel. The score of this method was in average for recall and precision of 62.94% and 87.30% (Figure 6.4d).

6.3.5 Knowledge-driven approach

The knowledge-driven method can be used as a post processing of the panel extraction. It validates or rejects panel candidates using the proposed model (Section 5.4). In the model, the only rule about the panel is that they should be contained in a image. The score of this method was in average for recall and precision of 61.88% and 87.42% (Figure 6.4e) using the independent panel extraction approach (Method I).



(a) Arai's method

(b) Ho's method

(c) Method S

(d) Method I

(e) Method K

Figure 6.4: Panel extraction score details for each image of the eBDtheque dataset (Appendix C).

6.3.6 Comparison and analysis

Table 6.1 compares the average results we obtained for the three proposed methods and two methods from the literature.

The main drawback of Arai's method is the binary segmentation with a fixed value that works best for comic books with a white background but it is not appropriate for digitization defaults and paper yellowing. This method is only appropriate for webcomics with a perfectly white background allowing a clear separation between panels (Figure 6.4a).

Ho's method is adaptive to background variation but considers as panel all the elements that are directly included (first topological level) in the page background.

Table 6.1: Panel extraction evaluation results.

	R (%)	P (%)	F (%)
Arai [8]	20.0	18.75	19.35
Ho [67]	49.76	68.74	57.48
Method S	**64.24**	83.81	72.72
Method I	62.94	87.30	**73.14**
Method K	61.88	**87.42**	72.46

This produces an over-segmentation by including some text which is inside the frameless panels or inside implicit balloons. When those out of panel elements are small, they prevent the connected panel separation process to work properly because the stopping criterion is based on element size.

The first method we proposed (Method S) extracts panel and text simultaneously and then other elements in a sequential manner. It assumes that there are distinguishable clusters of connected-component in the image which is not always the case. Despite the limitations, it is still more powerful than the two baseline methods from the literature, mainly because of the more generic pre-processing steps.

The second proposed panel extraction method (Method I), based on connected-component clustering, is simple to implement, fast and efficient for comics with disconnected panels (separated by a white gutter). This method is not appropriate for gutterless comics (e.g. some mangas) or strips without panel borders such as those with an extra frame around several panels. Another weakness is when panels are connected by other elements; they may not be split as desired but will remain clustered as panel anyway.

Method K benefits form the domain knowledge to filter out irrelevant panel candidates from the independent extraction approach (Method I). Nevertheless, the validation by the expert system was not significant here because the low level processing had already reached the limits of the model (Section 5.2.2).

6.4 Text extraction evaluation

In this section we evaluated our three propositions for text extraction and compare them to methods from the literature. We selected Arai's method as a baseline, as it is the most advanced method that has been applied to comic book images. All the evaluations of text localisation were performed on the 4691 text lines of the eBDtheque dataset [63] at object bounding box level (Section 6.2.1). Text recognition evaluations have been performed on the same dataset but using the accuracy metric presented in Section 6.2.3.

6.4.1 Arai's method

We implemented Arai's method [8] which is a sequential approach that requires panel and balloon extraction as input, as presented in Sections 6.3.1 and 6.6.1 respectively. This method was developed for grey-scale Japanese mangas and is divided in five steps: pre-processing, blob extraction, feature extraction, text blob selection and text blob extraction. Pre-processing consists in an adaptive bi-level segmentation at 30% (chosen empirically) above the average pixel value and mathematical morphology to group potential text letters into blocks (similar to the Run Length Smoothing Algorithm). Note that the threshold selection method can exceed the maximal pixel value and then produce an empty segmentation. The size of the kernel was not specified in the original paper, we used $3 * 5$ pixels. Blob extraction is performed by imposing a minimal size of the blob relatively to the size of the balloon in which they are included ($Balloon.width/20$ and $Balloon.hight/40$). The extracted features are the *average text blob width* and the horizontal coordinate of its centroid. These features are used to classify text / non-text blobs according to two rules based on blob inter-distances and alignments. Text blobs are then extracted from the original image and directly sent to an OCR system (Figure 6.5).

Figure 6.5: Sample of text extraction process. (a) Original balloon text image; (b) Threshold; (c) Morphology-Erosion filter; (d) Morphology-Opening filter; (e) Extracted blob from original image. Image from the original paper [8].

This approach was only developed for vertical text block extraction but we adapt it to also handle horizontal text. Its adaptation to horizontal text consists in switching width and height related parameters in the pre-processing and feature extraction steps. Also, the kernel used for mathematical morphology is rotated by 90°.

This method yields an average recall of 2.81% and precision of 1.63% (Figure 6.6a).

6.4.2 Sequential approach

The method proposed in Section 3.2 does not require any particular parameter, only the number of clusters has to be fixed to $k = 3$ according to the domain knowledge of comics (Section 3.2). Note that the method is designed to simultaneously extract panel regions which do not interfere with text. This method yields an average recall of 54.68% and a precision of 56.92% (Figure 6.6b).

6.4.3 Independent approach

The method presented in Section 4.3 consists in a bi-level segmentation, text/graphic separation, text line generation and finally text line recognition. A comparison between two threshold selection methods and two colour-to-grey image conversions are shown in Table 6.2.

Table 6.2: F-measure results for fixed ($th = 250$) and adaptive threshold selection method corresponding to combined (RGB to grey) and separated (Luminance layer) colour to grey conversion methods.

	Fixed	Otsu
RGB to grey	11.74	**51.35**
Luminance layer	11.66	51.31

Text and graphics separation is performed using the Mahalanobis' distance between the grey-level distribution of the whole image and each connected-component region (Section 4.3.3). A validation on the eBDtheque dataset showed that the best Mahalanobis' distance is $d_m = 2$. Text line recognition is used as last filtering operation to validate the presence of text inside the candidate regions, its average benefit is shown in Table 6.3. This method yields an average recall of 64.14% and precision of 70.28% (Figure 6.6c).

6.4.4 Knowledge-driven approach

The knowledge-driven method can be used as a post processing of the text extraction. It validates or rejects text candidates using the proposed model (Section 5.4). In the model, the only rule concerning text location is that it should be contained in a panel. Considering the best method for panel and text extraction method from Tables 6.1 and 6.3 respectively, this method yields an average recall of 39.99% and a precision of 64.88% (Figure 6.6d).

6.4.5 Comparison and analysis

Table 6.3 summarises the recall, precision and f-measure of the evaluated methods.

Arai's method is a sequential approach that requires panel and balloon regions as input. As shown in Table 6.6, balloon extraction is very low using this approach since the propagation of errors narrows down the text extraction (Table 6.3). Moreover, this text extraction method uses several thresholds which reduce its fields of application.

Method S improves substantially the recall thanks to its genericity, moreover, even if the panel extraction is performed at the same time, it does not bias the text extraction because text areas are extracted from the whole page and not from panel regions. Method I is slightly better using RGB to grey image conversion than the

Table 6.3: Text localisation results.

	R (%)	P (%)	F (%)
Arai [8]	2.81	1.63	2.07
Method S	54.91	57.15	56.01
Method I (without OCR)	**67.21**	41.54	51.35
Method I (with OCR)	64.14	**70.28**	**67.07**
Method K	39.99	64.88	49.48

luminance layer only (Table 6.2). It outperforms Method S in recall only when not using the last OCR filtering step and both recall and precision when using the OCR. The difference for Method I when using or not the OCR system is about an increase of 26.74% of precision and a loss 3.07% of recall. The validation by the expert system in Method K improved the precision as expected but also resulted in a drop in recall. The drop of Method K can be explained by the fact that the text extractor is also able to detect texts which are not in the speech balloons but the model considers them as noise (rejected). All the methods still encounter difficulties with certain types of text that can be found in the comics (e.g. illustrative text, graphic sounds) due to strong deformation or colour variations.

6.5 Text recognition evaluation

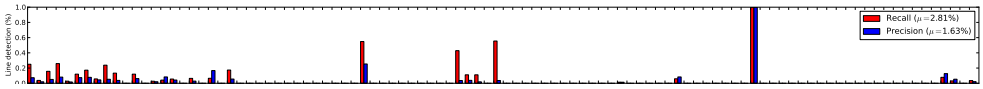
We evaluated text recognition using string edit distance between a predicted text recognition given by the OCR and its corresponding transcription in the ground truth (Section 6.2.3). The eBDtheque dataset is composed of English, Japanese and French texts. We evaluated the transcription given by the OCR engine Tesseract version 3.02.02 with the provided training data for each language[4]. This was performed at the text line level, taking as correct the text lines that were transcribed exactly as the ground truth transcription, ignoring case (lower/upper case) and accents for predicted and ground truth regions. Table 6.4 shows text recognition results given the text line position from the ground truth (best reachable score) and from the best automatic text localisation method (Method I in Table 6.3).

Table 6.4: Text recognition accuracy $A_{textReco}$ results from automatic text line localisation at an edit distance (ED) of 0, 1 and 2.

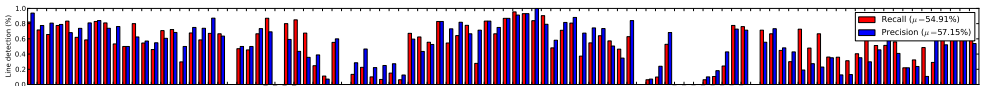
Text extraction method	$ED = 0$	$ED = 1$	$ED = 2$
Method I	11.11	16.53	20.67

We obtained a score of $A_{textReco} = 11.11\%$ for perfect transcription from the automatic text extraction and OCR method which constitutes a baseline for future work on text recognition on the eBDtheque dataset [63]. Table 6.4 shows the results

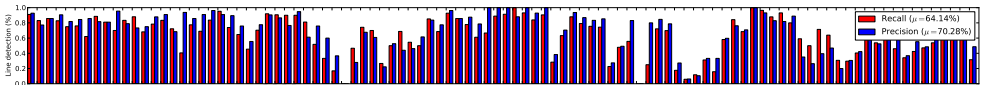
[4] `https://code.google.com/p/tesseract-ocr/downloads/list`



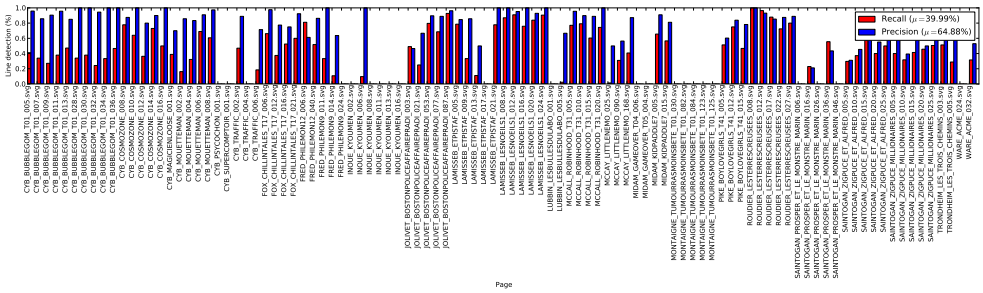
(a) Arai's method



(b) Method S



(c) Method I using the best bi-level combination method (rgb2grey + Otsu) and OCR



(d) Method K

Figure 6.6: Text line extraction score details for each image of the eBDtheque dataset (Appendix C).

for a more relaxed evaluation where we also considered as correct the text lines at a text edit distance equal to one and two, the accuracy rises up to 20.67%. Note that the average text line length is quite short in comic books compared to other documents, the distribution is given Figure 6.1.

6.6 Balloon extraction evaluation

In this section we evaluate our three balloon extraction approaches (Method S, I and K) and compare them to two methods from the literature. We compare our results to Arai [7] and Ho [67] that are two state of the art methods (Section 2.2.1). The evaluations are performed on the 1091 balloons of the eBDtheque dataset [63] at object and pixel levels in order to provide a performance evaluation for localisation and contour analysis purposes (Sections 6.2.1 and 6.2.2).

In the eBDtheque ground truth, 84.5% of the balloons are closed and 15.5% are not. We call the latter "implicit" balloons, as they are implicitly defined by other

cues in the image, and not by an explicit border (Section 3.3). Thus we do not expect to reach 100% recall and precision for the regular balloon extraction methods because they are not designed for implicit balloons. On the other hand, the implicit balloon extraction method is able to extract closed balloon as well.

6.6.1 Arai's method

Arai's method in a sequential method that extracts balloon inside panel regions in order to extract text inside [8]. It it similar to its panel extraction work-flow also presented in the same paper (Section 6.3.1), without image inversion which basically detect white blobs (assuming speech balloons have a white background). Not all the white blobs correspond to balloons in the comic book images, especially for the monochrome ones. The method selects blob candidates according to four rules based on blob size, white pixel occurrence, vertical straight line detection and width to length ratio. We re-implemented all the rules and heuristics of the original paper except for the straight line detection that we rotated of 90 degrees for images using horizontal text.

The score of this method was in average for recall and precision of 13.40% and 11.76% (Figure 6.9a).

6.6.2 Ho's method [67]

Ho's method is a sequential approach, similarly to Arai's approach, that extracts balloon inside panel regions in order to extract text inside [67]. First it converts the image from RBG to HSV colour representation and selects candidate regions with a high value (V) and low saturation (S) level. To reduce the number of candidate regions, a second selection is applied according to size and shape. Small regions are removed and only regions with a ratio between the number of pixel in the region over the number of pixel in its bounding higher than 60% are kept. In the original paper, only the ratio between the region area and its bounding box was given (60%). Low value (V) and high saturation (S) values were not mentioned in the original paper, we decided to combine them by removing the value (V) from the saturation (S) to compute a brightness level $B = S - V$ in order to reduce the number of parameters. We fixed $B = 200$, only region with a higher brightness level are kept. We considered as small regions the region with a width or height inferior to three pixels.

The score of this method was in average for recall and precision of 13.96% and 24.73% (Figure 6.9b).

6.6.3 Sequential approach

In Sections 3.3.1 and 3.3.2 we have presented two sequential approaches for regular and implicit balloon extractions from text line positions. Here, we evaluate them separately and detail the benefits of the different energies used for implicit balloon

extraction (active contour model). For both regular and implicit balloon extractions, the input text regions are given from the sequential text extraction method (Section 3.2) since it is the previous step of this approach.

Regular balloons Regular balloon extraction method is based on white blob selection according to the central position of text lines in the balloon. This central position is converted as a confidence value $C_{balloon}$ for each balloon candidate. The best score of $R = 37.25\%$ and $P = 45.19\%$ was obtained by considering only balloons with $C_{balloon} > 50\%$ (Figure 6.9c).

Implicit balloons The implicit balloon extraction method was evaluated using three different scenarios in order to highlight the benefits of both active contour theory and domain specific knowledge. First, the performance using active contour model with only the internal energy, then with the distance transform based external energy described in Section 3.3.2 and third, we added the second external energy from domain knowledge E_{text} presented in Section 3.3.2 (Table 6.5). The best average score for object level detection ($R = 57.76\%$ and $P = 41.62\%$) is detailed for each image of the dataset in Figure 6.9d.

Table 6.5: Implicit balloon performance evaluation at object and pixel levels using the original form and the proposed (with E_{text}) energy functions.

	Object level			Pixel level		
Energy function	R (%)	P (%)	F (%)	R (%)	P (%)	F (%)
$E = E_{int} + E_{ext}$	56.01	40.40	46.94	69.40	**83.98**	76.00
$E = E_{int} + E_{ext} + E_{text}$	**57.76**	**41.62**	**48.38**	**74.80**	82.67	**78.55**

6.6.4 Independent approach

As introduced in Section 4.4.1, Method I it is an independent approach that does not require any previous element extraction. This method requires one parameter which is the minimum number of children $minNbChildren$ of a balloon (number of included connected-components). We set $minNbChildren = 8$ according to the parameter validation illustrated Figure 6.7. The value 8 corresponds to the first peak in the distribution of the number of letters per speech balloon. Note that in Figure 6.7, there are about 3.5% of the balloons below the selected threshold that contain one or two letters, usually punctuation marks. We voluntary omitted them here in order to avoid over segmentation.

This regular balloon extraction method attributes a confidence value $C_{balloon}$ to each balloon candidate according to the alignment of the connected-components it contains (children), similarly to regular balloon extraction of Method S. The best score of $R = 52.68\%$ and $P = 44.17\%$ was obtained for $C_{balloon} > 50\%$. Note that we observed a drop in the performance when considering only balloons with a confidence

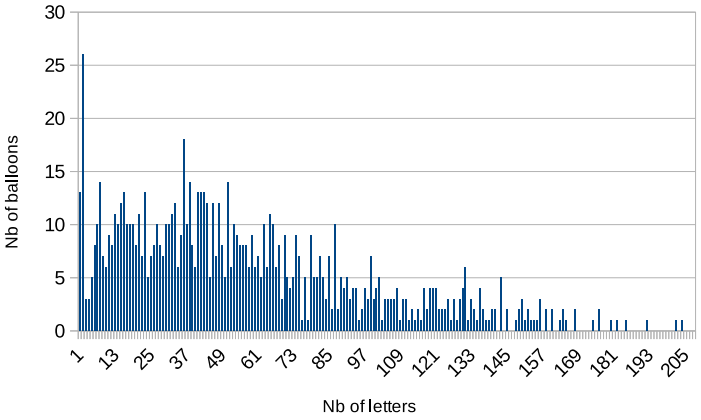


Figure 6.7: Distribution of the number of letter per speech balloon in the eBDtheque dataset [63].

value $C_{balloon}$ higher than 60%. Figure 6.9e confirms that our method works best for images where the balloons are closed, well segmented and with non-cursive text inside in order to create a sufficient number of connected-components.

6.6.5 Knowledge-driven approach

The knowledge-driven approach can be used as a post processing of the balloon extraction. It validates or rejects balloon candidates according to contextual information from the proposed model (Section 5.4). In the model, the rules concerning the balloons specify that a balloon should be contained in a panel and should not contain other balloons, panels or comic characters. Considering the best method of balloon extraction (Method S implicit in Table 6.6), the score of this method was in average for recall and precision of 39.56% and 48.65% (Figure 6.9f).

6.6.6 Comparison and analysis

Balloon localisation aims to evaluate how the presented methods are good at predicting where the balloons are in the image, it is performed at the level of object bounding box (Section 6.2.1). Table 6.6 compares the average results we obtained for the best performance of the three proposed methods and two methods from the literature.

Our methods outperforms [8], thanks to its genericity, since it can process all the image styles of the eBDtheque dataset. This was expected as [8] was specifically developed for mangas that have certain stylistic particularities.

Method A (regular) detects less than half of the balloons of this dataset (about

Table 6.6: Balloon localisation result summary.

	Object level			Pixel level		
Method	R (%)	P (%)	F (%)	R (%)	P (%)	F (%)
Arai [8]	13.40	11.76	12.53	18.70	23.14	20.69
Ho [67]	13.96	24.76	17.84	14.78	32.37	20.30
Method S (regular)	37.25	45.19	40.83	47.75	44.58	46.11
Method S (implicit)	**57.76**	41.62	**48.38**	**74.80**	**82.67**	**78.55**
Method I (regular)	52.68	44.17	48.05	69.81	32.83	44.66
Method K (regular)	39.55	**48.64**	43.63	56.65	41.89	48.17

37.25% recall). This approach is not originally designed for detecting implicit balloons but knowing the position of the panels allows closing open panels by drawing their bounding box. The bounding box creates an artificial frame that sometimes recover implicit balloons extraction (Figure 6.8).

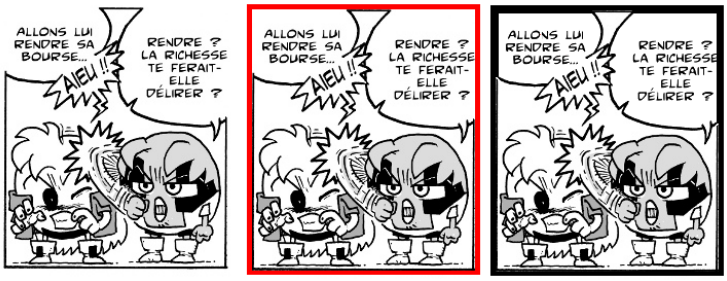


Figure 6.8: Implicit balloon recovery by closing open panels. From left to right: original panel, panel bounding box detection, panel with artificial frame from its bounding box that closes two balloons. The implicit balloons are now regular (closed).

However, it has two advantages in comparison with the proposed active contour based method. First, the precision of the balloon contour detection is maximal because it is a connected-components from the image (not contour approximation or bounding box). Second, it is faster to compute because the position of the connected-component is known, no need to compute an optimal position.

The presented implicit balloon extraction version of Method S highly depends on the active contour initialization success. In this study, we relied on text detection as we assume it is the most common feature that balloons include, while past experiments have shown that its accurate detection is feasible and stable. A side-effect of this choice is that balloons sometimes can not be detected because they contain non-text information (e.g. drawings, punctuation). We believe there is room for improvement of the different energy terms we used. For example, the Gradient Vector Flow proposed by Xu [200] can be used as well to compute the external energy, especially in

the case of missing data balloon boundaries (implicit balloons). On the other hand, the ground truth of implicit balloons is at best questionable, as the exact position of the balloon contour is quite subjective. A way to circumvent this problem could be to annotate the boundary in a flexible way in the ground truth. Our results using active contour with distance transform show a significant improvement compared to the regular balloon extraction, thanks to the active contour theory that detects much more balloons, both open or closed, than the connected-component based methods (Table 6.6).

Sequential approaches such as Arai, Ho and Method S suffer from the limitation of dependency between the different processing. For instance, in Method S the performance of our text extractor is 56.01% (Table 6.3) which is used as input for balloon extraction so the balloon extraction is inevitably affected. Also, such approaches search for balloon inside the panel region, as a consequence, these methods can not extract balloons that overlap several panels. This is easily handled by independent methods such as Method I because it does not consider panel positions and search for balloons in all the part of the image. The post validation by the expert system (Method K) reached the maximum precision but decreased the recall of the extraction which resulted in a drop of the overall f-measure by almost 5% compared to Method S. The drop in recall is due to the balloons that were correctly extracted but which contained undetected text regions (rejected because the proposed model validates only balloons that contain text regions).

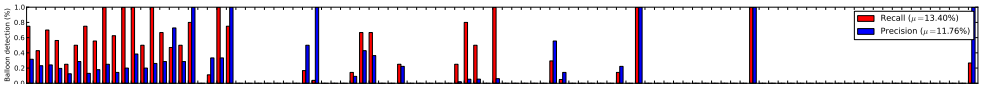
6.7 Balloon classification

From our knowledge, this is the first work that tackles speech balloon classification. We analyse the effects of the two features and the prior knowledge we presented in Section 4.4.2. We evaluate the proposed method using a subset of 914 balloons from the 1091 balloons present in the original eBDtheque ground truth (only "smooth", "wavy" and "spiky" balloons). The repartition of the balloon types is 87.96%, 8.75% and 3.29% of type "smooth", "wavy" and "spiky" respectively. This unbalanced repartition reflects the usual speech balloon types repartition in comics.

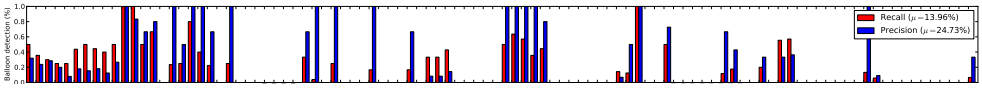
In the shape/contour separation process (Section 4.4.2), we first cast the original signal o into 360 values and then smooth it in order to approximate the shape signal s. The shape signal is also composed by 360 values where each of them corresponds to the local mean in a window of size M from the original signal (Equation 6.6). Note that in order to avoid side effects, we copy $M/2$ values from the opposite side (wrapping). A validation of parameter M have been performed and its best value was $M = 7$.

$$s(i) = 1/M \sum_{j=i-M/2}^{i+M/2} o(i) \tag{6.6}$$

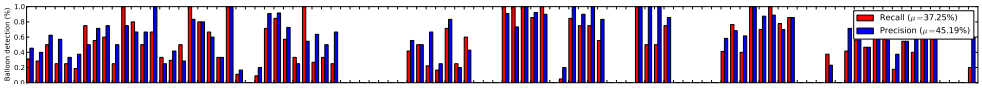
We evaluated our method based on a one-versus-all scenario. Classification results



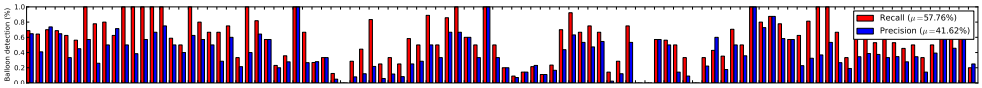
(a) Arai's method



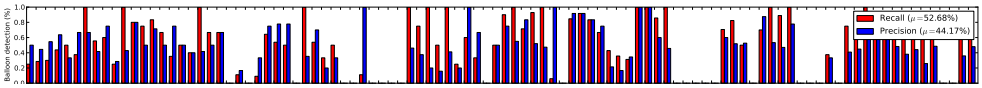
(b) Ho's method



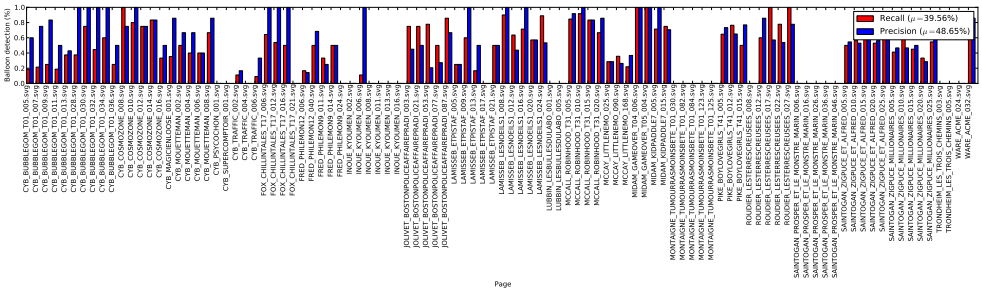
(c) Method S (regular)



(d) Method S (implicit)



(e) Method I



(f) Method K

Figure 6.9: Balloon extraction score details at object level for each image of the eBDtheque dataset (Appendix C).

are presented in Table 6.7 for different descriptors. First we show the effect of the two features f_1 and f_2 separately without *a priori* information $P(l_1) = P(l_2) = P(l_3)$ and then using the prior probabilities about the classes repartition from the eBDtheque dataset $P(l_1) = 0.88, P(l_2) = 0.09$ and $P(l_3) = 0.03$. We detailed our best results for balloon contour extraction (for $D = (f_1, f_2)$) in the confusion matrix Table 6.8.

As mentioned Section 4.4.2, the region of the tail may influence balloon classification. In order to measure this influence, we remove the tail by ignoring the points

of the contour that are between the two points defined as tail origin (Section 3.4), see Figure 6.10. Note that this process relies on the tail detection performance (Section 6.8).

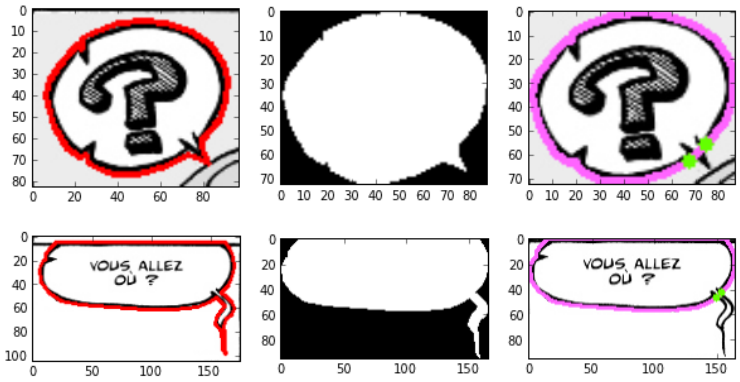


Figure 6.10: Two examples of tail removal for balloon classification. From left to right, original contour detection in red, corresponding mask and partial contour removal between the two green points that represent the tail origin. The final contour is represented in purple in the right figures.

Table 6.7: Classification result accuracy for different descriptor configuration. The global accuracy represents the number of good classification divided by the number of elements to classify independently from the class.

	Accuracy per class (%)				
	Smooth	Wavy	Spiky	Average	Global
$D = (f_1)$	22.51	68.75	33.33	41.53	26.91
$D = (f_2)$	79.73	36.25	70.00	61.99	75.60
$D = (f_1, f_2)$	76.62	**60.00**	**73.34**	**64.33**	75.05
$D = (f_1, f_2) + priors$	**98.26**	11.25	63.34	57.62	**89.50**

The different accuracy results Table 6.7 show that the first feature f_1 is more appropriate for separating "wavy" from the rest while the second feature f_2 is more efficient for "smooth" and "spiky" separation. Their combination (f_1, f_2) is therefore the best solution for dividing the balloons into these three classes. Adding the prior knowledge about the testing classes repartition improves the score of the "smooth" class and decrease the two others. This is due to the consideration of the probability of 88% to be in the "smooth" class by the Bayesian classifier. The confusion matrix Table 6.8 shows good classification results for the "spiky" class even for cases that are quite different in terms of drawing styles (Figure 6.11). Concerning the "smooth" and "wavy" classes, some speech balloons are very hard to classify out of context causing confusion (Figure 6.12). Also, the tail of the balloon creates some noise and the radius

Table 6.8: Confusion matrix for smooth (S), wavy (W) and spiky (Z) balloon contour classification results with $D = (f_1, f_2)$.

		Ground truth			Ground truth no tail		
		S	W	Z	S	W	Z
Prediction	S	**616**	180	8	114	349	341
	W	30	48	2	7	**53**	20
	Z	4	4	**22**	4	13	13

average ($\bar{o}$) we use to normalize the descriptor features may not be appropriate for elongated balloons.

Figure 6.11: Correct classification examples for "spiky" (top) and "smooth" classes (bottom).

6.7.1 Results analysis

As mentioned in Section 4.4.2, the balloon contour contains part of information that is also important to retrieve in the context of comic book understanding (e.g. tail, tone information). In this context, the location of the balloon is not sufficient. Such analysis requires a finer extraction of the balloon contour at pixel level (Table 6.6).

The proposed method covers the comics balloon classification in general, except for open balloons. However, in this particular case, we could use extra features (e.g. tail type recognition, language analysis) to get more information about the text tone. Also, other distortion measures have to be investigated, especially for "wavy" contours that are the most difficult contour to classify using the proposed method. Concerning the tail region, its removal is not has significant as expected, it only improves the accuracy of "wavy" balloon classification (Table 6.8). This is due to the reduction of the number of points of the balloon contour to 360 values which reduce the impact of the tail region.

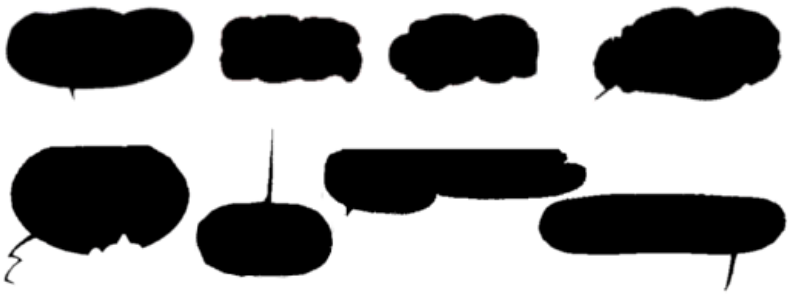


Figure 6.12: Wrong classification examples which have been classified as "smooth" instead of "wavy" in the first row and "wavy" instead of "smooth" in the second row.

6.8 Tail detection evaluation

In this section we evaluate the tail extraction method on the 1091 balloons of the eBDtheque dataset [63] from three balloon extraction methods, one manual and two automatic extractions. The automatic extraction methods are the two best methods for regular (Method I) and implicit (Method S) from the summary given in Table 6.6. The manual method is the balloon contours from the eBDtheque dataset. The comparison of the three balloon extraction method will highlight the part of error which comes from the balloon extraction and the proper tail extraction method's error when there is not possible error from the balloon extraction. As introduced in Section 6.2.4, we use two accuracy metrics, one for the predicted position of the tail tip $A_{tailTip}$ and the other one for the tail direction $A_{tailDir}$ (Table 6.9).

Table 6.9: Tail tip position ($A_{tailTip}$) and tail direction ($A_{tailDir}$) extraction results from automatic and manual balloon contour extractions.

Balloon extraction method	$A_{tailTip}$ (%)	$A_{tailDir}$ (%)
Ground truth	**96.61**	**80.40**
Automatic (regular balloons)	55.77	27.59
Automatic (regular + implicit balloons)	64.34	20.21

6.8.1 Result analysis

The proposed method is very good at locating the tail tip when it exists and also at giving a confidence value equal to zero when the balloon has no tail. The few mistakes concerning the tail tip position detection is due to very small tails or to other contour variations similar to the tail which confuse the vertex selection process (especially in the case of tail-less balloons). The tail direction is sensitive to the quality of the pixel-level balloon contour extraction in the area of the tail tip. For instance, in Figure 6.13, the tail positions are represented by a green line on the balloon contour

and tail directions are written in blue (from the eight cardinal coordinates). Left hand side balloons have not been detected without tail because of the contour variations, a tail have been mistakenly predicted. For the right hand side balloons, the tail tip position have been correctly predicted but not the tail direction because of a strong change of direction of the last segment. The directions have been predicted according the white balloon background region and not from the balloon boundary that indicates a different direction (NE and SE instead of S and E).

The tail tip position and tail direction highly rely on the performance of the balloon contour extraction. They are satisfactory when balloon contours are perfectly extracted (manual in Table 6.9) but decreases quickly according to the automatic balloon extractor method performance. The importance of a precise pixel-level segmentation in the region of the tail tip is illustrated by the difference of 7.38% for $A_{tailDir}$ between the regular and implicit automatic methods due to the lack of precision of the active contour extraction method.

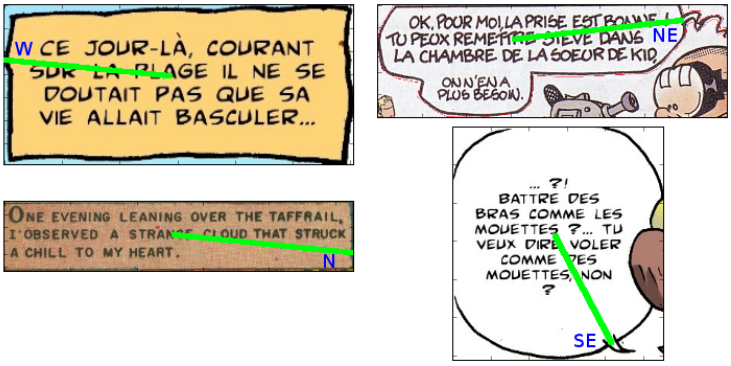


Figure 6.13: Examples of difficult balloons for tail tip position (left column) and tail direction (right column) extractions.

6.9 Comic character extraction evaluation

We evaluated the detection of comic characters on the 1597 characters of the eBDtheque dataset. In the eBDtheque ground truth, only 829 (52%) of the character instances are speaking (and 48% are not). Therefore, we do not expect to reach 100% recall and precision with Methods S and K because they are only able to extract speaking characters. The independent Method I is potentially able to retrieve all the character instances of colourful comics, given one example for each of them. This condition requires a lot a manual work thus we selected only 22 different characters which represent 352 instances in total.

Similarly to the other object extraction methods, the overlapping ratio between a predicted character region and a region from the ground truth should be above 50% ($a_0 > 0.5$) for the predicted region to be considered as valid (Section 6.2.1). This

ratio can be relaxed according to the application (e.g. face only, full body, cropped body). An example of the impact of this ratio on validation and rejection of predicted regions is shown on Figure 6.14. For instance, the predictions that are presented in the panels of Figure 6.14 could be accepted if the target was to roughly detect comic character locations. In this case, all the predicted regions would have been accepted as correct if the validation criterion was set at $a_0 > 0.1$.

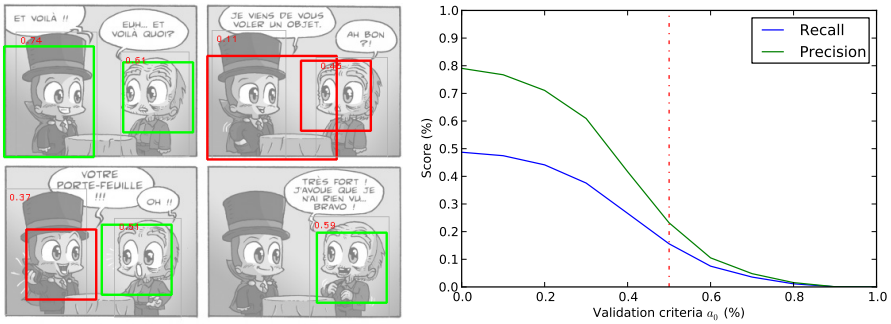


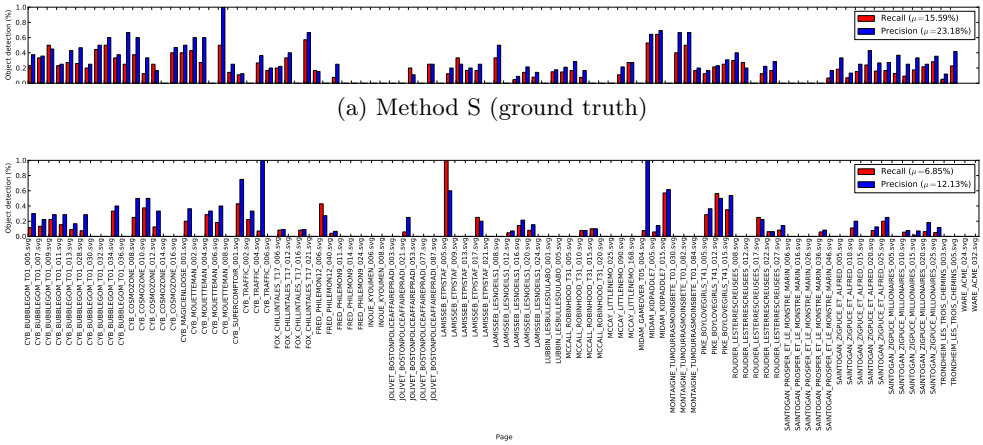
Figure 6.14: Example of predicted region considered as true (green) or false (red) positives according to the validation criteria $a_0 > 0.5$ (dashed lines in the right hand side figure). Thin grey rectangles are the ground truth regions and the red numbers in the top left corner of each ground truth region is the value of the overlapping ratio a_0. Image credits: [39].

6.9.1 Sequential approach

Method S is a sequential approach that computes a region of interest for comic characters from the tail tip position and tail direction inside a panel region (Section 3.5). Its position is shifted in the region of the tail tip according to the tail direction which is quantized in eight cones of $\pi/4$ radius. We evaluated the performance of this method given the input information of panel, balloon and tail from the eBDtheque ground truth and from the best automatic extraction methods from Table 6.1, 6.6 and 6.9 respectively (Figure 6.15a and 6.15b). Also, two sets of images have been tested, one including all the comic character instances of the eBDtheque dataset and another one with only the speaking characters, as this method is only able to predict speaking character location from the tail direction (Table 6.11).

6.9.2 Independent approach

Method I consists in an independent comic character instance spotting given an example. To perform this experiments, we asked the users to manually highlight the object position as described in Section 4.5.2. The selection usually includes hair, head and top body colours because lower body parts are often hidden by the frame border or the posture. From the user pixel selection, we computed the bounding box to show



(a) Method S (ground truth)

(b) Method S (automatic)

Figure 6.15: Character extraction score details at object level for each image of the eBDtheque dataset for speaking and non-speaking characters (Appendix C).

that the system does not need a precise object colour selection. We fixed the descriptor size to $N = 3$ according to the evaluation shown on Figure 6.16. In the retrieval process, we set $T = N$ which means that the confidence of the candidate region equal 100% (all the query colours must be present in the candidate region to be considered as a correct detection). We used three squared window sizes according to the query height: half of the query height, query height and twice the query height.

The following results were obtained with the common evaluation measures of recall, precision at object level which means that we considered as correctly detected the objects that are overlapped by a detected region without considering the percentage of overlapping. Recall (R) is the number of correctly detected object divided by the number of objects to detect. Precision (P) is the number of correctly detected objects divided by the number of detected objects. We consider as candidate regions, regions that have $C = 100\%$ confidence (containing at least all the descriptor colours).

We evaluated our method on a subset of 34 comic book images from 10 different albums which represent 22 comic characters that appear 352 times in total. This subset has been chosen from relevant images from the eBDtheque dataset that are coloured and with a sufficient amount of character instances. We obtained 90.3% recall and 46.7% precision in average. Results are detailed in Table 6.10 where album 1 to 10 correspond to the following image identifiers in the eBDtheque dataset: MIDAM GAMEOVER T05, TARZAN COCHON, CYB COSMOZONE, CYB MAGICIENLOOSE, CYB MOUETTEMAN, LAMISSEB LESNOEILS1, LUBBIN LESBULLESDULABO, MIDAM KIDPADDLE7, PIKE BOYLOVEGIRLS T41 and TRONDHEIM LES TROIS CHEMINS 005 (Appendix C).

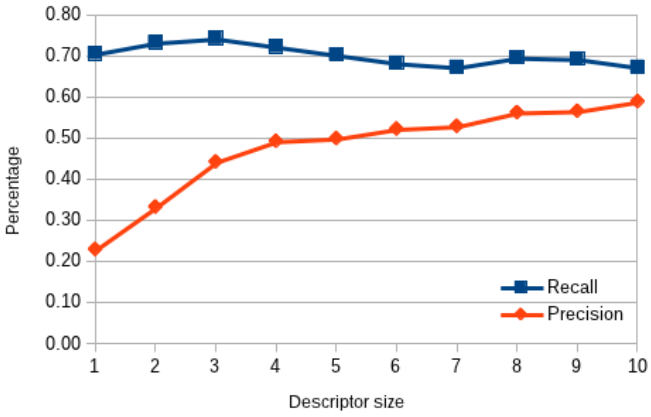


Figure 6.16: Descriptor size validation. Measure of recall (blue square) and precision (red diamond) detection for different descriptor sizes from a testing set of 17 comics pages. The maximum recall is for $N = 3$.

6.9.3 Knowledge-driven approach

The knowledge-driven method can be used as a post processing of the comic character extraction using only the properties of the comics model or an extra processing if we consider ROI refinement and non-speaking character spotting as illustrated Section 5.4. Here we use only the validation part of the expert system without ROI refinement and extra character spotting because these last approaches have not been completed during this thesis work. The expert system validates or rejects character candidates according to contextual information from the proposed model (Section 5.2.2). In the model, the rules concerning comic book characters are that they should not contain any other extracted region because we assume that a character can not contain panels, balloons, text or other comic characters (Section 5.4). If they do, they are automatically deleted. We post processed the output of the automatic extractions from the sequential approach (Method S) and we obtained an average score for recall and precision of 3.59%, 12.68% (Figure 6.17).

6.9.4 Comparison and analysis

The comic characters extraction are presented in Table 6.11 for Method S and K. Method I is evaluated separately on a smaller subset of the dataset. Note that the scores in Table 6.11 are given for the whole dataset including the characters that are not speaking.

This low level of performance comes from various reasons. First, the limited capacity of the extractor to process speaking characters. When we focused on the subset

Table 6.10: Character detection performance

Album	Nb pages	Nb character	ID	Recall	Precision
1	10	2	0	92.3%	16.8%
			1	91.8%	65.0%
2	1	1	2	75.0%	26.5%
3	5	4	3	100.0%	50.0%
			4	93.3%	60.9%
			5	100.0%	33.3%
			6	100.0%	100.0%
4	1	1	7	85.7%	85.7%
5	4	2	8	100.0%	75.0%
			9	50.0%	27.8%
6	5	2	10	71.8%	10.1%
			11	100.0%	29.9%
7	2	2	12	100.0%	13.8%
			13	100.0%	54.5%
			14	100.0%	55.6%
8	2	2	15	81.8%	75.0%
			16	100.0%	12.7%
9	2	1	17	83.3%	58.8%
10	2	4	18	81.3%	35.1%
			19	100.0%	48.4%
			20	92.9%	17.1%
			21	90.9%	24.4%
10	**34**	**22**	**22**	**90.3%**	**46.7%**

of 829 speaking characters, the recall and precision increased by 14.45% and 4.85% respectively, for panel and balloon regions from the ground truth (GT). Second, the variability of character styles in the eBDtheque dataset; third, the error propagation of previous processes (panel, text, balloon, and tail extractions) that are required to guess comic character locations (differences between "GT" and "auto" results in Table 6.11).

As mentioned Section 6.9.2, Method I have been evaluated on a subset of the eBDtheque dataset composed by coloured images and with a sufficient amount of character instances. The experiments shown interesting detection results based on a user-defined colours query descriptor. Detection result examples are illustrated Figure 6.18.

The detected region aims to localize the smallest regions containing all the colours

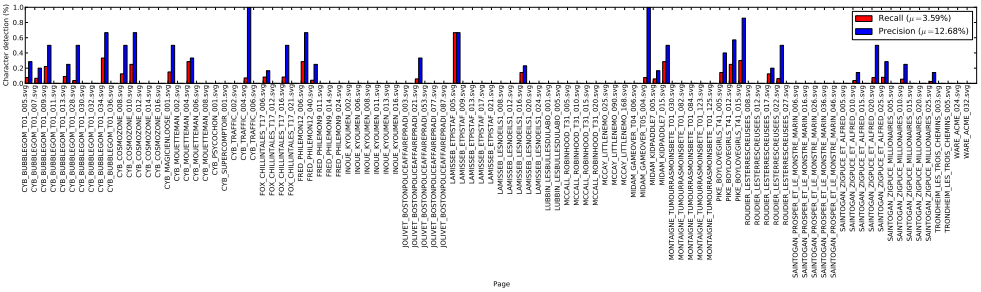


Figure 6.17: Character extraction score details using Method K for each image of the eBDtheque dataset (Appendix C).

Table 6.11: Comic character localisation result for method S from ground truth (GT) and automatic (auto) panel and balloon element extractions.

	Speaking only			All characters		
Methods	R (%)	P (%)	F (%)	R (%)	P (%)	F (%)
Method S (GT)	**30.04**	**28.03**	**29.00**	**15.59**	**23.18**	**18.64**
Method S (auto)	11.77	11.82	11.79	6.84	12.13	8.75
Method K (auto)	6.15	12.30	8.20	3.59	12.68	5.60

of the query descriptor. This region is usually smaller than the ground truth region which is defined at character bounding box level. A post-processing is needed to compute the character bounding box from the detected region by extending it to the colour region boundaries.

Correct detections show the variety of comics character position and deformation that we are able to detect with the presented framework. There are few missing detections and over detections are essentially due to other comics character detection and image pre-processing (colour quantization).

The proposed framework gives promising results for contemporary comics, we also tried on historical one such as FRED_PHILEMON12 and MCCALL_ROBINHOOD from the eBDtheque dataset (Appendix C) but the printing process generates a lot of noise (thickness of ink). A pre-processing smoothing is required in this case [88, 201]. The colour palette we used has 256 colours, we believe that we can improve the presented method by computing the palette from the user's query in order to ignore unwanted colours and speed up the process. The sliding windows approach does not allow to segment comic characters at pixel level, a distance measure between the colour mask regions could solve this. Also, once we retrieved all colour-similar objects, we could learn a deformable shape model and try to find more objects occurrences based on shape information in a second stage. These proposals will be the subject of future research topics.

Figure 6.18: Each line shows a character query region bounding box in white (left column), a correct detection (middle column) and a wrong detection (right column). Green rectangles represent the ground truth region and the red rectangle corresponds to the region detections. The corresponding character IDs in Table 6.10 are 12, 8, 0, 10, 16 and 2 from top to bottom.

6.10 Knowledge-driven analysis overall evaluation

In this section we evaluate the proposed knowledge-driven analysis system as a whole. Conversely, in the previous four sub-sections we evaluated and compared each low level content extraction separately. Here we take advantage of the independent content extraction method (Method I) and the knowledge base proposed in Chapter 5,

to evaluate the performance of the expert as the whole, which includes visual and semantic elements. First we evaluated the two models and then the proposed sequence presented in Section 5.4.

6.10.1 Comics model evaluation

Considering the lack of accurate numbers on the overlapping ratio of comic book' elements in the literature, we estimated statistically what this *significant proportion* would be using the eBDtheque dataset [63]. Figure 6.19 shows the percentage of panels, balloons, text lines and characters that fit the proposed model (Section 5.3.2, as a function of their covered area.

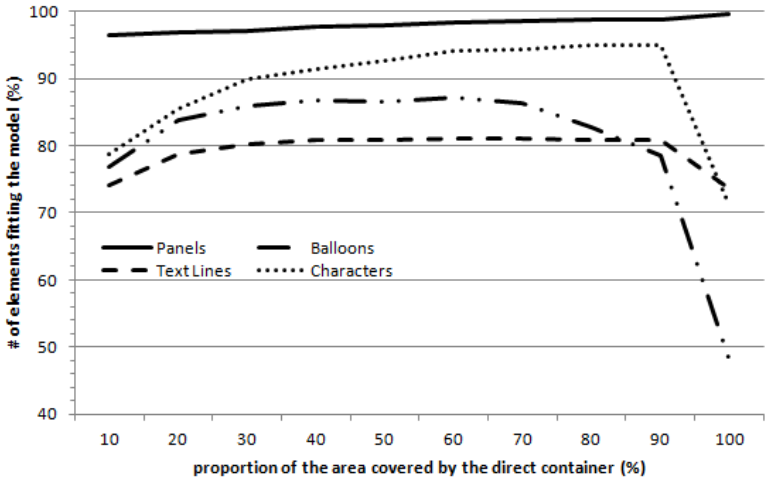


Figure 6.19: Percentage of panels, balloons, text lines and characters from the eBDtheque dataset 6.1 that fit the definition given in 5.3.2 according to the area proportion that need to be covered.

Considering ideal proportion covered for each type of element, we obtained the following overall scores: 99.6% of the panels, 87.4% of the balloons, 81.6% of the text lines and 94.9% of the characters are in accordance with the assumed constraints in the eBDtheque dataset.

6.10.2 Framework evaluation

We evaluated our framework during the two iterations of the process loop introduced in Section 5.4 and particularly at the beginning (Step 1: hypothesis) and at the end (Step 3: inference) of each iteration. Low level content extraction such as panels P, balloons B, text T and comic characters C have been evaluated separately in the four previous subsections, here we combine them throughout the two iterations of the process loop. Two high level element retrieval processes are also evaluated here, the

semantic link between the spoken text and speech balloons $STSB$, and between the speech balloons and speakers $SBSC$.

We evaluate our framework using the F-measure of panel P, balloon B, text T and character C extractions. Also, the accuracy of the two semantic links $STSB$ and $SBSC$ is measured (Section 6.2.5). Results are presented in Table 6.12 and the change in the amount of information discovered throughout the process is illustrated in Figure 6.20.

Table 6.12: F-measure score evolution for panel (P), balloon (B), text (T) and characters (C) throughout the two iterations. Scores for the semantic links $STSB$ and $SBSC$ are given as accuracy.

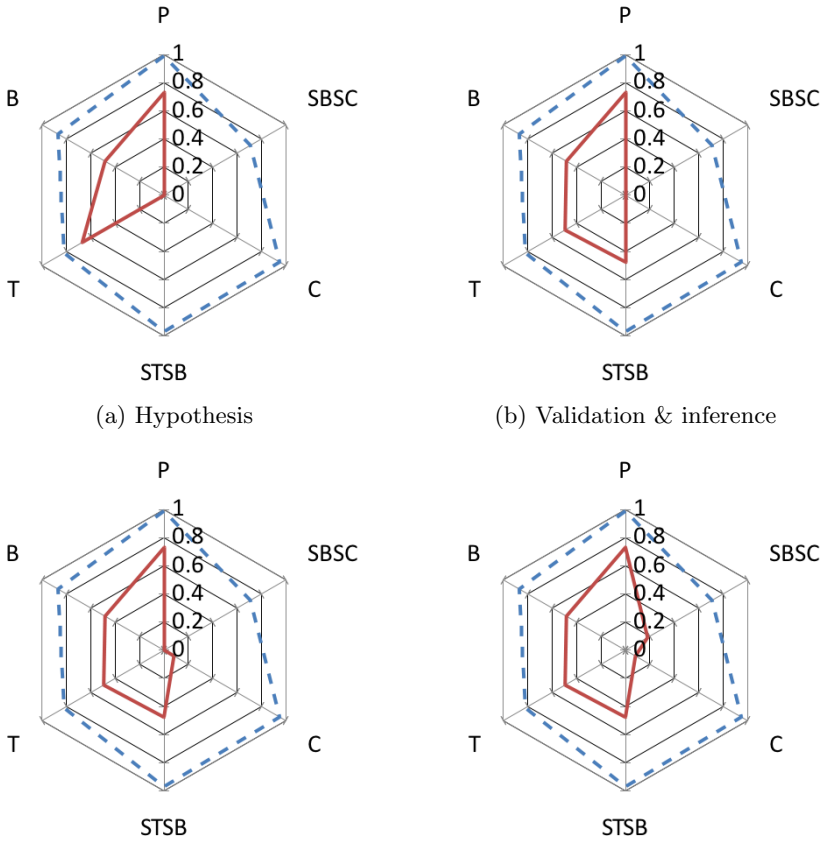
Step	P	B	T	C	$STSB$	$SBSC$
1 - Hypothesis	73.14	48.38	67.07	0	0	0
1 - Validation & inference	72.46	43.63	49.48	0	47.61	0
2 - Hypothesis	72.46	43.63	49.48	11.79	47.61	0
2 - Validation & inference	72.46	43.63	49.48	8.20	47.61	18.23

Process details Figure 6.20 shows the evolution of the performance of the knowledge-driven system after the first and second iteration of analysis (Section 5.4) over the eBDtheque dataset [63]. We also represented the maximum reachable limit using the proposed model (Section 6.10.1) using a dashed line in the same figure.

The performance of the first iteration is measured after the initial extraction of simple elements which are considered as hypotheses by the expert system (Figure 6.20a) and after validation by the expert system (Figure 6.20b). Between the first initialization (1 - Hypothesis) and validation (1 - Validation), the F-measure remains stable for P (less than 1% difference) and decreases by 4.75% for the balloon B and 17.59% for the text T while more than 47% of the semantic links between speech text and speech balloon ($STSB$) are retrieved. The global decrease is related to the actual behaviour of the expert system which is only able to filter out elements. Nevertheless, remaining elements are consistent and allow semantic relation inferences. The expert system applies the inference rules to automatically specify the balloons that have a tail and their included text as speech balloon SB and speech text ST respectively to create the semantic links $STSB$ between each pairs.

At this point, no comic book characters are discovered because their rules in the ontology are related to elements that are not yet discovered.

The newly inferred links are processed by the expert system along with previously validated regions during a second iteration in order to get more information by trying to apply more rules from the knowledge base. This time, the expert system can make use of rules related to characters because the speech balloons are now part of the knowledge base. Panels and speech balloons are used to create hypotheses for the location of characters and then the low level processing locates them more precisely within these regions (Figure 6.20c). Finally, the expert system validates the newly



(a) Hypothesis (b) Validation & inference

(c) Hypothesis (d) Validation & inference

Figure 6.20: Change in performance for panels P, balloons B, text lines T, comic characters C and the semantic links $STSB$ and $SBSC$ at the hypothesis and evaluation steps during the two loops of the process. The dashed blue line represents the best score using our model on the data extracted from the ground truth data (optimal condition). The solid red line is the performance of the framework using all the automatic extractions and semantic link inferences.

discovered regions and infers the semantic links between speech balloons and speaking characters $SBSC$ (Figure 6.20d).

Low level processing The best extraction performance is obtained for the panels that are usually the easiest elements to extract from a page. The lowest extraction performance is for the comic characters C. There are various reasons for this. First, the limitation of the extractor to process speaking characters; second, the variability of character styles in the eBDtheque dataset; third, the error propagation of previous

processes (panel, text, balloon, tail and semantic link extractions) that are required to guess comic character locations.

Semantic links The expert system is able to retrieve 47.6% of the $STSB$ and 18% of the $SBSC$ relations, which represents more than 25% of what could possibly be detected using the proposed model. It should be stressed that these numbers represent the efficiency of the last process of the whole framework pipeline. Individual errors at each recognition and validation step of the pipeline are propagated to the final semantic association between elements ($SBSC$). Therefore a single improvement in the detection or the validation of any kind of element would have an impact at the semantic association level.

Note that given the panel, balloon and character position from the ground truth, the accuracy of the expert system in predicting the semantic relations is about $A_{STSB} = 96.9\%$ and $A_{SBSC} = 70.66\%$. The 3.1% of missed *isLineOf* relations came from balloons that are not compliant with our model. In the same way, among the 829 *isSaidBy* relations, that link speech balloons to speaking characters, 9.5% are undetectable because they are generated from balloons outside panels. This comparison clearly demonstrates that the expert system is very good at predicting semantic links as soon as the low level extractions are efficient.

6.11 Conclusions

In this chapter we have first introduced the dataset and the metrics we used. Then, we evaluated each of the contributions for content and semantic element retrieval and compared them to other methods from the literature. Finally, proposed models combined with independent content extraction methods have been evaluated.

We presented the construction of the eBDtheque dataset, the first comic image dataset and associated ground truth of comic book images, containing spatial semantic and bibliographic annotations. This dataset is now publicly available for the scientific community. The corpus can easily be extended to enlarge the diversity of the dataset and provide consecutive pages or full albums in order to allow a wider level of comic books analysis. New semantic annotations can be added, such as the view angle and the shot type for panels, more text characteristics, comic character's role, profile and relationships.

The sequential approach (Approach S) is an intuitive approach like most of the previously proposed methods in the literature. It efficiency is maximal when all the image content is present and structured such as "classical" comic books. It gives the best average and precision score for text, balloon and character extraction which demonstrate that linking different content extractions at low level processing is relevant for comic book image analysis.

The independent approach (Approach I) has similar performance results as the previous approach only for panel extraction. It becomes more efficient than the pre-

vious approach when processing unusual comic image (e.g. no panel, no text or no balloon). However, it is slightly less powerful than sequential approach in average on the eBDtheque dataset because this dataset is mainly composed by classical comics. Nevertheless, this approach can easily handle particular issues as it does not require previous element extraction.

The knowledge-driven approach (Approach K) is able to improve the overall precision but not the recall in the presented configuration because it only uses rejection constraints. The strength of this approach is that it uses a standard and exchangeable description of the image content, from visual to semantic level, that can be easily be enriched or adapted to specific or large collections. The addition of conversion and creation constraint types (not only rejection) can give the ability to this intelligent system to recover missing information.

Processing times are quite short for the sequential and independent approaches but the knowledge-driven approach requires reasoning operations that are time consuming. Panel and balloon extractions can run in real time on a regular laptop or mobile devices. Text extraction process highly depends on the OCR speed. Comic character region of interest are almost computed instantaneously (calculus of maximal rectangle coordinates) but character spotting is very slow due to colour quantization and multi-scale sliding window processes.

Chapter 7

Conclusions

Throughout the dissertation, several methods for comic book image analysis and understanding have been presented. This chapter summarizes each main chapter by revisiting their contributions, strengths and weaknesses. Finally, an overview of the future research possibilities in the area of comics analysis and understanding is discussed.

7.1 Summary and contributions

In this thesis we have presented three different approaches for comic book image analysis. Chapter 1 has summarized the evolution of sequential art from its creation to the 21st century with the impact of the Internet, its market place and the growing interest of the investigation of this research field.

In Chapter 2, an overview of the state-of-the-art methods has been presented. Here we have handed the methods in their context of document analysis and then highlighted the challenges of comic book document analysis due to their specific design process. All the past studies related to comic book image analysis have been reviewed into four different categories for panel, balloon, text and comic character respectively. Literature reviews in each of these categories have been presented along with the advantages, disadvantages in different scenarios. Also, state-of-the-art methods for holistic understanding of document images have been reviewed and a list of the most advanced existing applications has been given.

Then, in Chapter 3, we have introduced a first approach of sequential analysis of comic book image content by extracting elements using their relations one after others in order to guide the retrieval process. The major contribution in this work is to take advantage of previous extracted elements for predicting the region of interest of speaking characters. The main motivation of this work came from the idea of an intuitive approach for retrieving simple elements in order to facilitate the retrieval

process of more complex ones.

Chapter 4 has presented a independent analysis technique based on the separation of each extraction process. The main contribution was the introduction of a text extraction method which is independent from balloon locations and able to detect out-of-balloon text regions such as illustrations, page numbers, captions, author names and some sound effects. The benefit of standard OCR systems was also stressed. The next contributions of this work were balloon contour type classification and comic character spotting based on partial user-defined example.

In the Chapter 5, we have introduced a knowledge-driven analysis system for comic book images. The main contribution was the introduction of two models, one for general comics and another one for comics-related image processing domain knowledge. The two models are queried by an inference engine to guide independent image processing, retrieve the contents and its interactions in order to provide a high level of description (including contextual and semantic information). There are several limitations of the independent extraction approach presented in Chapter 4 that can be recovered using these models. This contribution solves the limitation of error propagation which is implicit to the sequential approach presented in Chapter 3, while using relations between elements defined by the domain knowledge.

Finally in Chapter 6, we have provided an experimental evaluation of all the proposed approaches against state-of-the-art methods or different scenarios. Moreover, advantages and disadvantages of each method have been pointed out.

In general, in this thesis we have proposed different contributions to improve previous works from the literature as for panel, text and regular balloon extraction. Other contributions are first studies such as implicit balloon extraction, balloon type classification, tail detection and semantic analysis. Detailed experimentations were performed to evaluate each contribution compared to some state-of-the-art algorithms. In order to compare objectively the methods, a first dataset and ground truth have been built and provided to the community.

7.2 Future perspectives

There are ideas that have come out during this thesis work but could not be explored in the three year time. Some of the future perspectives of this thesis are listed below.

Panel extraction is the more studied part of comics analysis so far and still require some efforts for implicit panel detection (e.g. when the panel border is partially or not drawn) and connected panels (e.g. several panels connected by an overlapping element). Gutter-free panel extraction methods also require more experiments to confirm their robustness against a huge variety of comics but public datasets are lacking.

Text extraction is an essential challenge that has been tackle only partially. Current methods focus on speech text and further efforts has to be made to detect other types of text such as illustrative text (e.g. brand name, storefront), caption, page

number, author signature, and sound effects (onomatopoeia). The next step will be text recognition which is an open issue as comics can be handwritten or typewritten with various specific fonts. Semi-automatic font learning methods are probably one of the solutions to overcome the lack of performance of standard OCR systems.

Balloon extraction shown good performances when combined with text extraction but requires more efforts when processed independently, in order to extract balloons that do not contain text as well (e.g. symbol, drawing, punctuation). Implicit balloons are important to consider as well but it is at best questionable as the exact position of the balloon contour is quite subjective. Balloon classification is at his early stage, more balloon types have to be investigated and multi-segment classification would probably be more accurate for balloons with non-homogeneous contour variations (e.g. implicit, adjacent to other elements). Speech balloon classification can also be improved by analysing the nature of the contained text using natural language processing (NLP). Balloon classification should be completed by a semantic analysis to give a meaning to each class given a particular album (e.g. smooth for dialogues, wavy for thoughts) and finally characterize the contained text.

The first work about tail detection presented in this dissertation gave promising results that allow one-to-one connections between speech balloons and speaking characters but multi connections have to be considered as well (e.g a speech balloon with two tails pointing to two different characters saying the same thing).

Comic character localization and identification are still an open issue for both supervised and unsupervised methods due to the variability of character styles and postures. Nevertheless, combining colour and multi-part shape description seems to be a good way to follow in order to describe the characters for spotting purposes. Unsupervised comic characters localization and identification require a high level description method able to take into account heterogeneous elements in the image (e.g. ontologies, graphs).

Finally, we published the first public dataset to initiate collaborative research but more data are needed to be fully representative of the diversity of comics, to train algorithms and to evaluate research works related to comics understanding.

Appendix A

Pre-processing

Throughout this thesis work we have used several pre-processing methods to extract image content. The typical pre-processing tasks are noise removal, segmentation, enhancement, perspective correction, page curl removal, skew correction (text) and skeletonization. In this appendix we detail some of the segmentation methods which are related to comic book image analysis.

A.1 Segmentation

Image segmentation is a division or separation of the image into regions of similar attribute. The most basic attribute is image luminance amplitude for monochrome image or colour components for a colour image. Image edge and texture are also useful attributes for segmentation [149].

A.1.1 Region-growing

Region-growing methods are region-based segmentation algorithms that mainly rely on the assumption that the neighbouring pixels within one region have similar values. The common procedure is to compare one pixel with its neighbours. If a similarity criterion is satisfied, the pixel can be set as belonging to the cluster as one or more of its neighbours. The selection of the similarity criterion is significant and the results are influenced by noise in all instances[1].

One of the most famous region-growing segmentation is the seeded region-growing method [2,120]. This approach examines neighbouring pixels of initial seed points and determines whether the pixel neighbours should be added to the region and so on.

[1] http://en.wikipedia.org/wiki/Image_segmentation#Region-growing_methods

The results of the method highly rely on the selection of the seed points. From the seed points, we grow a 4 or 8-connected neighbourhood for the pixels adjacent relationship. Different parameters can be added, such as the minimum area threshold that will define the minimal region area returned by the algorithm. The similarity measure is the criterion which should be verified to accept a neighbouring pixel in the current region [57]. The segmentation by watershed, developed with mathematical morphology, belong to this category as well.

A.1.2 Split and merge

Split-and-merge segmentation is a region-based segmentation based on a quadtree split of the image. This method starts from the whole image (root of the tree) and if it is non-uniform (not homogeneous pixel repartition), then it is split into four son-squares (the splitting process), and so on so forth. Conversely, if four sub-squares are homogeneous, they can be merged as several connected components (the merging process). The node in the tree is a squared or set of squared regions. This process continues recursively until no further splits or merges are possible [69, 84].

A.1.3 Contour-based

This approach wants to take advantage of the fact that there is a detectable transition between two related areas. The oldest methods use edge detector such as the well know Canny filter, to highlight the pixels that appear to belong to a contour [22]. The reconstruction of closed contours is often difficult. See Appendix B.1 for closed contour extraction.

Irregular and partial contour segmentation have been studied as well, using deformable models that are more permissive. The most famous model have been proposed by Kass, it is called the active contour model or snake [83]. A curve $\mathbf{v}(s) = [x(s), y(s)], s \in [0, 1]$ is initialized in the region of the contour to detect and it moves through the spatial domain of an image to minimize the energy function until perfectly fitting the image region (Equation A.1).

$$E = \int_0^1 \frac{1}{2} \left(\alpha \left| \mathbf{v}'(s) \right|^2 + \beta \left| \mathbf{v}''(s) \right|^2 \right) + E_{ext}(\mathbf{v}(s)) ds \tag{A.1}$$

where α and β are weighting parameters that respectively control the snake's tension and rigidity, and $\mathbf{v}'$ and $\mathbf{v}''$ denote the first and second derivatives of $\mathbf{v}(s)$ with respect to s. This functional energy is also called E_{int} for internal energy. The external energy function E_{ext} is computed from the image so that it takes on its smaller values at the features of interest, such as boundaries [200]. One of the proposed energy functions by Kass [83] is Equation A.2 which attracts the contour to edges with large image gradients.

$$E_{ext} = -|\nabla \mathbf{I}(x, y)^2| \tag{A.2}$$

A.1.4 Bi-level grey-scale thresholding

A survey counts forty thresholding methods from histogram shape-based to spatial and local methods [160].

The most famous method is Otsu's method [137] that assumes that a single channel image is dividable in two main clusters: the background and the foreground. It calculates the optimum threshold separating the two clusters so that their intra-cluster variances are minimal in the image histogram. Applying the Otsu's threshold selection to a grey-level image returns a binary image where foreground pixel are separated from background pixels.

The major problem with thresholding is that it considers only the intensity, not any relationships between the pixels (no spatial information). Another problem with global thresholding approaches is that changes in illumination across the scene may cause some parts to be brighter and some parts darker in ways that have nothing to do with the foreground and background objects (e.g. light and shadow variation). Those issues have been partially handled by local approaches determining thresholds locally instead of the whole image.

A.1.5 Multi-level colour thresholding

A first approach for multi-level thresholding is to apply bi-level thresholding to each of the colour layer in one or several colour space and merge them to give a single output. Two more natural approaches are divisive clustering and agglomerative clustering. In divisive clustering, the entire image is regarded as a cluster, and then clusters are recursively split to yield a good clustering. In agglomerative clustering, each pixel or region is regarded as a cluster and then clusters are recursively merged to yield a good clustering. There are two major issues in thinking about clustering:

- What is a good inter-cluster distance?
- How many clusters are there?

Addressing both issue automatically still an open issue if we don't have strong knowledge about the image content. Common clustering algorithm like k-means partitions the data (pixels values) into k sets by minimizing the within-cluster sum of squares [144]. It does not require any probabilistic reasoning or modelling [13]. K-means clustering aims to partition $(x_1, x_2, ..., x_n)$ observations into k clusters $(k <= n)$ $S = \{S_1, S_2, ..., S_k\}$ in which each observation belongs to the cluster with the nearest mean. This optimisation problem is formulated as formula A.3.

$$argmin \sum_{i=1}^{k} \sum_{n_j \in S_i} ||x_j - \mu_i||^2 \tag{A.3}$$

where μ_i is the mean of points in the cluster S_i.

Appendix B

Feature extraction

In this appendix we recall the principle of one of the most famous feature extraction approach used for stroke-based documents: connected-component labelling.

B.1 Connected-component labelling

Connected-component (blob) labelling scans all pixels from an image and groups them into components based on their connectivity [180]. All pixels in a so called "connected-component" share similar intensity values and are in some way connected with each other (part of the same component). Connected component labelling was original designed for binary and grey-level images with different measures of connectivity, but it can also operate on colour image [44]. Blob extraction is generally performed on a binary image from a thresholding step (Chapter A). Then, blobs can be described, filtered and tracked for many applications. It is a two pass algorithm that process rows and then columns. The first pass assigns temporary labels and record equivalences whereas the second pass replaces each temporary label by the smallest label of its equivalence class that gives an already existing or new label to each pixel according to their 4 or 8 neighbourhood connectivity.

Connected-component labelling can also be useful to retrieve the topology of the image content [178]. See outermost contour extraction for panel detection in Section 4.2.

Appendix C

Dataset

This appendix shows an overview of the eBDtheque dataset images and their categories. The corresponding ground truth annotations are described in the next Appendix D.

C.1 Image overview

Figure C.1 to C.5 show thumbnails of the eBDtheque dataset images.

C.2 Image categories

Table C.6 represents the composition of the eBDtheque dataset grouped by categories. Corresponding images to the identifiers are available in the previous section and on the dataset website [1].

[1] http://ebdtheque.univ-lr.fr/database/?overview=1

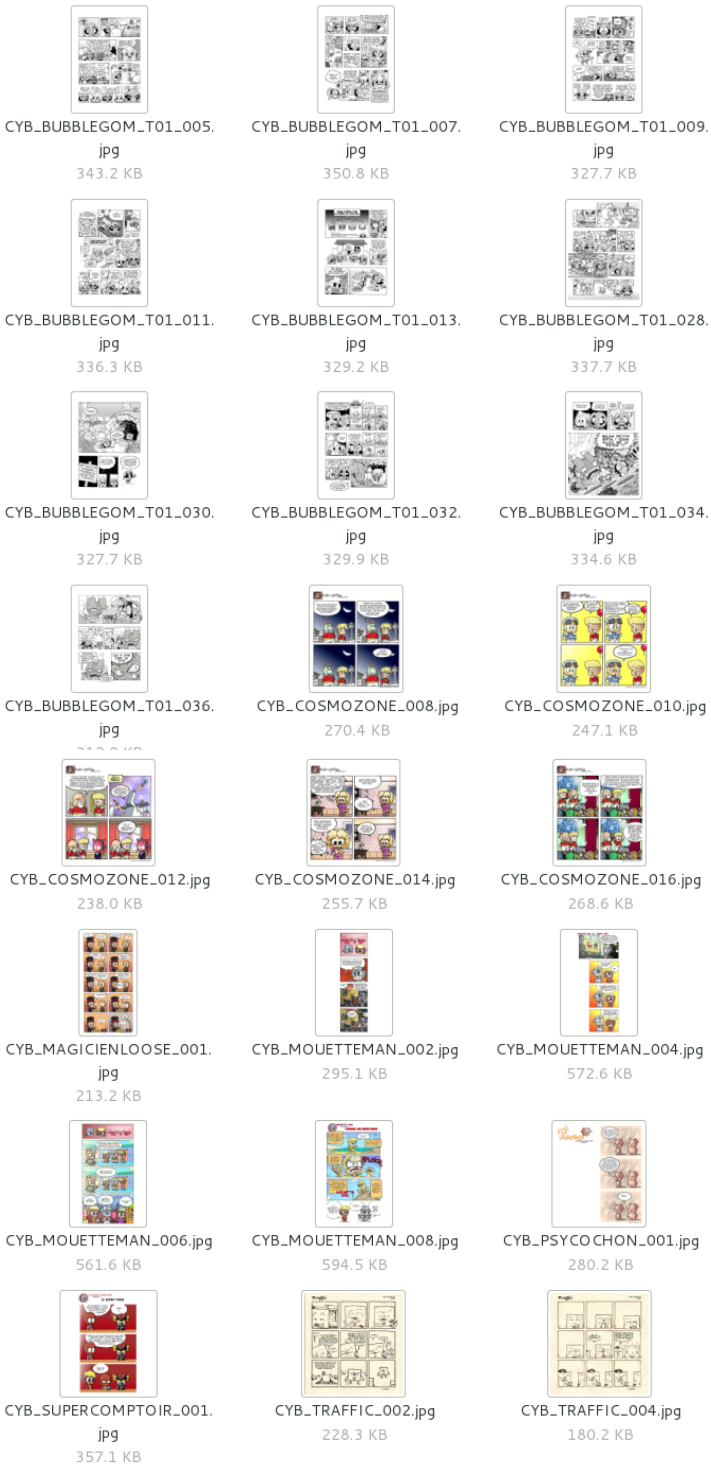


Figure C.1: Image group one.

Figure C.2: Image group two

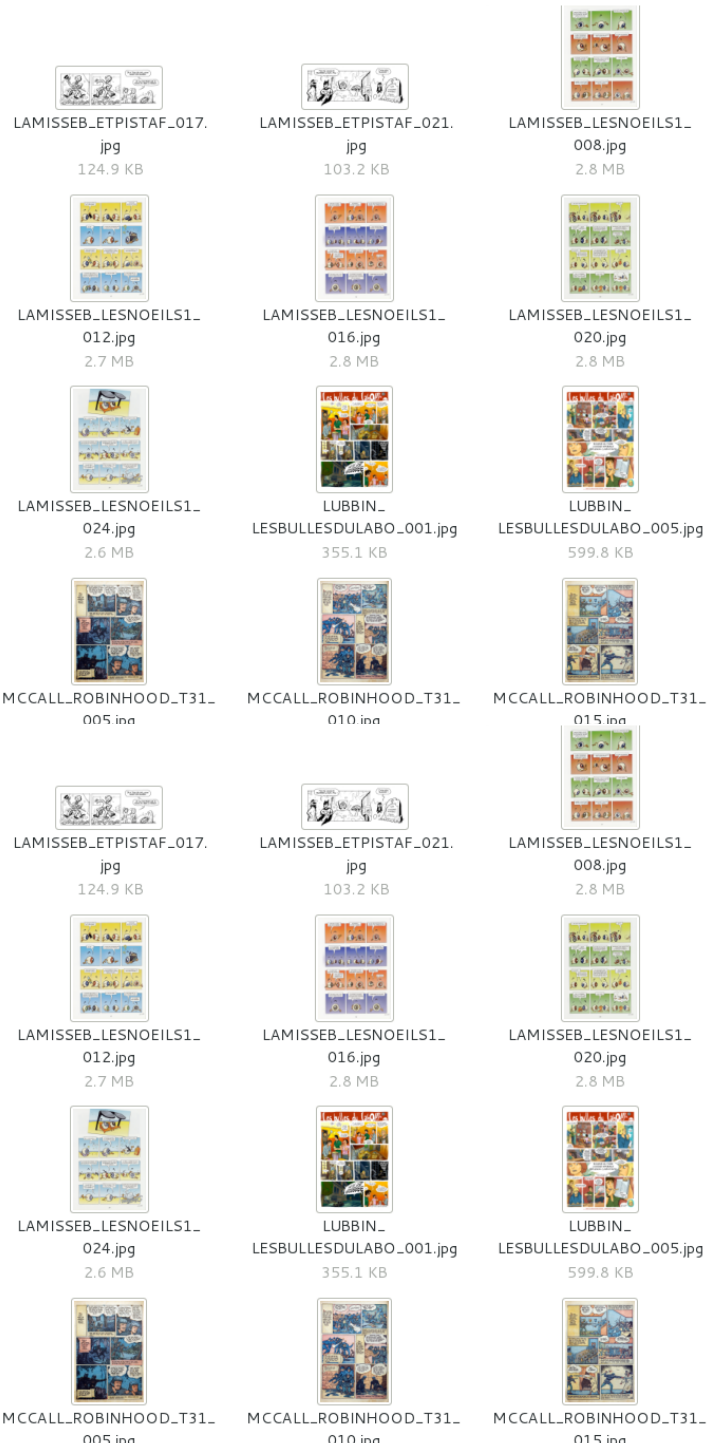


Figure C.3: Image group three.

Figure C.4: Image group four.

Figure C.5: Image group five.

IDs	Date	Category	Colour	Nb
CYB_BUBBLEGOM	2009	French webcomics	Grey	10
CYB_COSMOZONE	2008	French webcomics	Colour	5
CYB_MAGICIENLOOSE	2011	French webcomics	Colour	1
CYB_MOUETTEMAN	2011	French webcomics	Colour	4
CYB_PSYCOCHON	2010	French webcomics	Colour	1
CYB_SUPERCOMPTOIR	2011	French webcomics	Colour	1
CYB_TRAFFIC	2009	French webcomics	B&W	3
FOX_CHILLINTALES	1953	American comics	Colour	4
FRED_PHILEMON	2004	French comics	Colour	5
INOUE_KYOUMEN	2008	Japanese manga	B&W	6
JOLIVET_BOSTONPOLICE	2010	French comics	Colour	2
JOLIVET_BOSTONPOLICE	2010	French comics	B&W	3
LAMISSEB_ETPISTAF	2012	French webcomics	B&W	5
LAMISSEB_LESNOEILS1	2011	French comics	Colour	5
LUBBIN_LESBULLESDULABO	2012	French webcomics	Colour	2
MCCALL_ROBINHOOD	1946	American comics	Colour	4
MCCAY_LITTLENEMO	1905	French comics	B&W	1
MCCAY_LITTLENEMO	1905	French comics	Colour	2
MIDAM_GAMEOVER	2009	French comics	Colour	2
MIDAM_KIDPADDLE	2001	French comics	Colour	2
MONTAIGNE_ TUMOURRASMOINS	2011	French webcomics	Colour	5
PIKE_BOYLOVEGIRLS	1953	American comics	Colour	3
ROUDIER_ LESTERRESCREUSEES	2011	French comics	Colour	5
SAINTOGAN_PROSPER	1934	French comics	Colour	5
SAINTOGAN_ZIGPUCE_ ET_ALFRED	1952	French comics	Colour	5
SAINTOGAN_ZIGPUCE_ MILLIONAIRES	1928	French comics	Colour	5
TRONDHEIM_ LES_TROIS_CHEMINS	2000	French comics	Colour	2
WARE_ACME	2005	American comics	Colour	2

Figure C.6: Categories of the images in the eBDtheque dataset. "IDs" column corresponds to the beginning of the image identifier (file name and title of the corresponding grand truth file), "Date" is the release date and "Nb" the number of image of the category.

Appendix D

Ground truth

This appendix details the ground truth construction process, the visual (regions) and semantic information annotation, the file structure, the annotation quality assessment and finally the terms of use. This work is the result of an extensive collaboration with Clément Guérin [62] and many meetings with the eBDtheque project team of the L3i lab.

D.1 Ground truth construction

The ground truth has been defined in accordance to existing formalism in order to fulfil the needs of a large amount of researchers related to comics material. It integrates low and high level information such as spatial position of the elements in the image, their semantic links and also bibliographic information.

The comics art is extremely heterogeneous and our dataset voluntarily integrates albums that can be classified as unconventional. This leaves room for interpretation on the form which increase annotation variations by different people and decrease the uniformity of the ground truth. This precision level is used in several, widely used datasets [51, 204].

In order to cover a wide range of possible research matters, it has been decided to extract three different types of objects from the corpus: text lines, balloons and panels. We decided to do this first ground truth by drawing horizontal bounding boxes as close as possible from the feature and including all its pixels. We chose this level of granularity in order to limit the subjectiveness of the person making the annotation.

Hereafter we detail the two levels of annotation (visual and semantic) that form the ground truth and how they are indexed in a file. The combination of visual and semantic annotation provides the advantage of making this ground truth relevant for document analysis and semantic evaluation which are both part of the comics

understanding process.

D.1.1 Visual annotation

The first annotation consists in defining the spatial region where elements are located in the image. We describe here how the visual annotations have been performed for the panels, balloons, texts and comic characters.

Panels The frame or panels are defined as an image area, generally rectangular, representing a single scene in the story. There is always at least one panel per page; the entire page region can be used as panel if necessary. When a panel has a black border, the bounding box is placed as close as possible to its frame. Sometimes, images have not been scanned perfectly horizontally, it is then impossible to have an horizontal bounding box sticking exactly to the border. When the panel border is partially absent or suggested by the neighbourhood, the bounding box just defines the content of the panel. In all cases, the other elements (balloon, text, drawings) extending from the frame are truncated (Figure D.1).

Figure D.1: Example of three panel annotations. The bounding box (transparent red) is defined without taking into account of non-panel elements in all cases.

Balloons We define a balloon (phylactery or bubble) as the region of an image including one or several lines of text, graphically defined by an identifiable physical boundary or suggested by the presence of an arrow pointing to the speaker (the tail). Although rare, empty balloons (not containing lines of text) are also annotated if they are clearly identifiable by their shape or the tail representation. Pixel level annotation follows the contour of the balloon (Figure D.2c), while bounding box annotation does not consider the contour of phylactery and truncates the tail. Sometimes it crosses the entire panel and generates an unrepresentative position of the desired balloon (Figure D.2a). When the balloon is not closed (e.g. open contour) the annotated contour has to stick as close as possible to the contained text (Figure D.2b). Note that the first version of the ground truth (2013) has been defined at bounding box level ignoring the tail and the second version (2014) at pixel level following the contour variations and the tail region.

a) b) c)

Figure D.2: Example of balloon clipping: a) using bounding box excluding the tail, b) bounding box of non-closed balloon, c) pixel level contour annotation.

Text lines The text lines are defined as a sequence of text characters aligned in the same direction (Figure D.3a). This definition encompasses both speech text and narrative text, often located inside balloon, onomatopoeia (graphic sound) that are written or drawn directly in the panel without particular container. Comics are static graphics, the expression of emotions of a comic character is the joint action of drawing and text, sometimes in the form of a single punctuation symbol. For instance, an exclamation mark for surprise or a question mark for a misunderstanding. These isolated symbols convey information and are segmented as text line as well (Figure D.3b). Similarly, we have chosen to include in this category the illustrative text, such as a road sign or storefront (Figure D.3c). Although at the boundary between text and graphic, these elements are still invariably read by the reader and their annotation is potentially interesting for multiple purposes, including story and scene analysis.

Comic characters The comic character positions have been included in the second version of the ground truth only (2014). The concept of "character" may have different interpretations when used for comics and must be specified. Characters in a comic have not necessarily a human-like, or even living being appearance. Even so, it would be appropriate to annotate every character instance appearing in a box while some are only part of the scenery. Therefore, we have chosen to limit the annotation to the comic characters that emits at least one speech balloon in the album (minimal impact in the story). Their bounding box has been defined to maximize the region occupied by the comic character inside the box region. Therefore, some parts of the character such as arms or legs, are clipped sometimes (Figure D.4).

(a) Speech text into speech balloon

(b) Unique symbol (punctuation)

(c) Illustrative text (non-horizontal)

Figure D.3: Examples of text line position annotations.

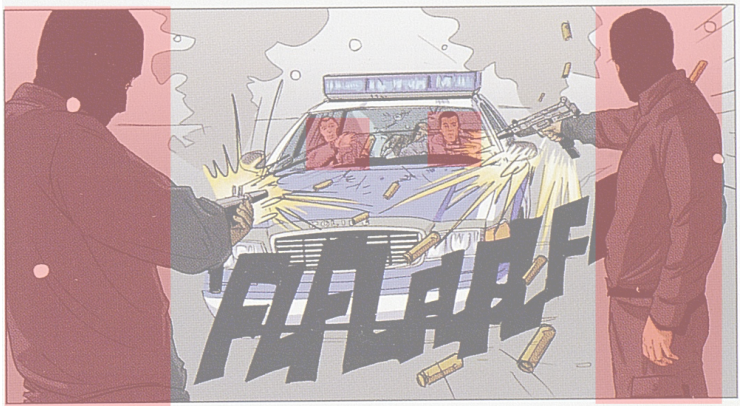


Figure D.4: Example of comic character annotation: sniper's arms are not included in the bounding box in order to maximize the region occupied by the sniper in its bounding box. The two snipers and the two characters in the car are annotated because they emit a speech balloon in a different panel. Image credits: [77].

D.1.2 Semantic annotation

This second level of annotation completes each spatial region with additional information about its semantic. Also, the image itself is annotated with extra information about its origin (e.g. album, collection, author and publisher).

Images The images, often assimilated as pages, has been annotated with bibliographical information, so that anyone using this ground truth is free to get his own paper copy of the comic books for extra uses. The first annotation is the page number (`pageNumber`) then the comic book title, from which the page has been picked up, and its release date (`albumTitle, releaseDate`), the series it belongs to (`collectionTitle`), the authors and editor names (`writerName, drawerName, editorName`) and, finally, the website and/or ISBN (`website, ISBN`). The album title is not mandatory for webcomics. Structural information about the page content has been added as well, such as resolution (`resolution`), reading direction (`readingDirection`), main language of the text (`language`) and single or double page information (`doublePage`).

Panels The panels are annotated with a `rank` metadata which stands for its position in the reading sequence. The first panel to be read on a given page has its rank property set to 1, while the last one is set to n, where n is the number of panels in the page.

Balloons Balloons are also annotated with a `rank` property that defines their reading order relatively to the image because balloons are not always included in panels. For a page containing m balloons, the first balloon's rank will be 1 and the last will be m. A second information concerns the `shape` of the balloon. This feature conveys information about how the contained text is spoken (tone). The type of `shape` is given from the following list {`smooth, wavy, spiky, suggested, other`} as pictured in Figure D.5. Finally, the tail tip position (extremity) and its pointing direction have been added into the second version of the ground truth. There are given through the `tailTip` and `tailDirection` properties. The possible values of the direction are reduced to the eight cardinal directions plus a ninth additional value for the lack of tail: {`N, NE, E, SE, S, SW, W, NW, none`}. In the second version of the ground truth (2014), we added the identifier of the comic character which is emitting the balloon `idCharacter`.

Text lines The text lines are associated with their transcription and the identifier of the corresponding balloon is added in `idBalloon` if the text line is included in a balloon. In a second step, the function of each text line was specified through metadata `textType`. We identified six distinct categories of text. The text that is used to tell the story and that can be either spoken (`speech`) or thought (`thought`) or narrated (`narrative`). On the other hand, there are textual information that bring

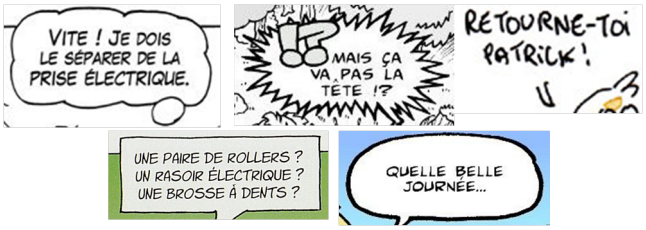


Figure D.5: Balloon contour styles, from top-left to bottom-right: cloud, spiky, suggested and two smoothed contour type. Note that the two last ones are not labelled as rectangle or oval because here we annotated the type of contour and not shape (Section 4.4.2)

timely and contextual information, they are the onomatopoeia (`onomatopoeia`) and drawn text (`illustrative`) as part of the drawing such as license plate, storefront, brand. The sixth category has been defined as notes (`note`) for embedded text in the page, such as the signature of the author, the page number, the title (every other readable text type).

Comic characters The comic characters are identified by `idCharacter` in order to be easily referred. Currently, it is the only metadata concerning the comic characters but it can be extended with information about character name, gender, posture, relationship etc.

D.1.3 File structure

The ground truth file structure have been thought according to existing comics related formalism such as Comics Markup Language (ComicsML) [118], Comic Book Markup Language (CBML) [194], Periodical Comics[1], A Comics Ontology [153], Advanced Comic Book Format (ACBF)[2] and the Grand Comics Database (GCD)[3]. See the Ph.D. thesis of Guérin [62] for an extended review.

As we wanted to keep the ground truth file system simple and easy to share, visual and semantic annotations about a given page are gathered in a single full-text file following the specifications of Scalable Vector Graphics (SVG). Besides being an open-standard developed by the World Wide Web Consortium (W3C) since 1999, the SVG format fulfils two essential needs for this database.

First, using a recent Internet browser or your favourite image viewer, it provides a simple, fast and elegant way to display the visual annotation of any desired object over

[1] `http://www.w3.org/wiki/WebSchemas/PeriodicalsComics`
[2] `https://launchpad.net/acbf`
[3] `http://www.comics.org`

a comic book page using layers. No need to install software such as Matlab, Adobe Illustrator or equivalent open source to visualize the ground truth information.

It is an XML-based vector image format that allows to display and interact with the annotated region, stored as polygon object in the SVG file, as desired using the Cascading Style Sheets (CSS) properties (Figure D.6).

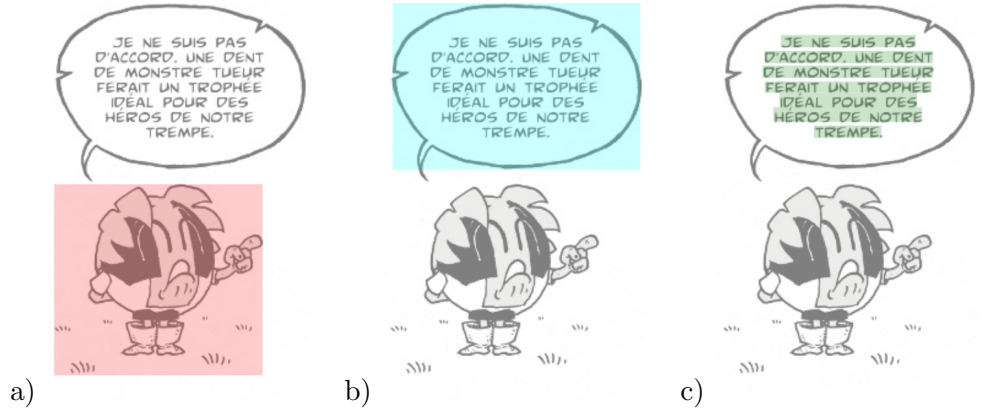


Figure D.6: Example of rendering for each class of element. For example, red for panels (a), cyan for balloons (b) and green for text (c). The opacity is set to 50% to allow seeing the corresponding image by transparency.

Each layer can be displayed or not in order to enhance the clearness of the annotations when browsing the database. Secondly, SVG being a XLM-based language, it makes the integration of semantic annotation very easy via the use of the predefined `metadata` element.

One ground truth file contains the complete description of one comics image. There is no hierarchical link between pages from a same comic book. Following the basic XML encoding information, a SVG file starts with a root `<svg>` element containing the title of the document, `<title>`, and five `<svg>` children with different class attributes. These contain annotations collected on five types of elements which are the page, panels, balloons, text lines and comic characters. The type of element in a tag is specified by its `class` attribute. The first tag, `class = ''Page''` contains description on the image and has two daughters. The first one, `image` has several attributes which specifies a link to the image file in the dataset `xlink: href` and two specifying the `width` and `height` of the image. The second, `metadata`, contains bibliographic information about the album and page properties described in Section D.1.2. The four following `<svg>` siblings, `<svg class=''Panel''>`, `<svg class=''Balloon''>`, `<svg class=''Line''>` and `<svg class=''Character''>` respectively contain the annotations about panels, balloons, text lines and comic characters. They all contain SVG `<polygon>` elements with a list of five or more points in a `point` attribute that define the position of the bounding box corners or the pixel-level contour. Note that the last point

is always equal the first one to "close" the polygon according to the SVG specifications. Those points are used by the viewer to draw polygons with the corresponding CSS style. Each <polygon> has a <metadata> child to store information on the corresponding polygon, according to the attributes list described Section D.1.2.

An example of ground truth file content is given Listing D.1.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<svg>
  <title>CYB_BUBBLEGOM_T01_005</title>
  <svg class="Page">
    <image
      x="0"
      y="0"
      width="750"
      height="1060"
      href="CYB_BUBBLEGOM_T01_005.jpg"
    />
    <metadata
      collectionTitle="Bubblegom_Gom"
      editorName="Studio_Cyborga"
      doublePage="false"
      website="http://bubblegom.over-blog.com"
      albumTitle="La_Legende_des_Yaouanks"
      drawerName="Cyborg_07"
      language="french"
      resolution="300"
      ISBN="979-10-90655-01-0"
      readingDirection="leftToRight"
      writerName="Cyborg_07"
      releaseDate="2009"
      pageNumber="5"
    />
  </svg>
  <svg class="Panel">
    <polygon points="53,95 268,95 268,292 53,292 53,95">
      <metadata
        idPanel="P01"
        rank="1"
      />
    </polygon>
    ...
  </svg>
  <svg class="Balloon">
    <polygon points="61,103 143,103 143,172 61,172 61,103">
      <metadata
        idBalloon="B01"
        shape="smooth"
        tailTaip="153,167"
        tailDirection="SE"
        rank="1"
```

```
    />
   </polygon>
   ...
  </svg>
  <svg class="Line">
   <polygon points="373,121 432,121 432,132 373,132 373,121">
    <metadata
     idLine="L01"
     idBalloon="B01"
    >
     LIKE YOU.
    </metadata>
   </polygon>
   ...
  </svg>
  <svg class="Character">
   <polygon points="84,153 261,153 261,298 84,298 84,153"/>
    <metadata idCharacter="C01"/>
   ...
  </svg>
  <svg class="LinkSBSC">
   <polygon points="34,234 56,235 79,340 79,339 34,234"/>
    <metadata
     idLinkSBSC="LSBSC01"
     idBalloon="B01"
     idCharacter="C01"
    />
   ...
  </svg>
</svg>
```

Listing D.1: Example of ground truth information stored in a SVG file

D.2 Ground truth quality assessment

When several persons are involved in the creation of a graphical ground truth, it is very difficult to obtain a perfectly homogeneous segmentation. Indeed, it could vary from one person to another because each person has a different sensitivity at reading comics and at integrating instructions. Therefore, in addition to the package of pages he was in charge of, each participant has been asked to annotate the panels of an extra page. This extra page was the same for everybody and was chosen for its graphical components heterogeneity. It contained ten panels from which, four were full-framed, five half-framed and one was frameless. We defined an acceptable error for the position of a corner given by several persons. The images of dataset being of different definitions, using a percentage of the page size makes more sense than using a specific number of pixels. We set this percentage p at 0.5% of the page height

and width in x and y. Given the definition of the test image of 750x1060 pixels, this makes a delta of +/- 5 pixels in y axis and +/- 4 pixels in x axis.

We asked to each one of the twenty involved persons to draw the four points bounding box of the panels ignoring text area. A mean position from the twenty different values has been calculated for each of them. Then, the distance of each point to its mean value is computed. Figure D.7 shows the amount of corners for a distance, centred on zero.

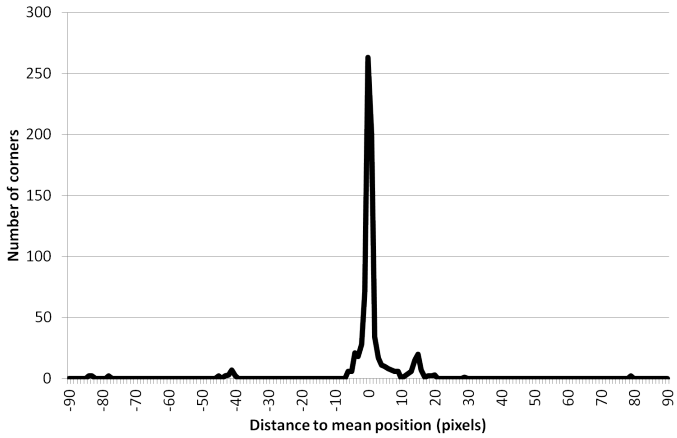


Figure D.7: Number of corners for a given standard deviation value. This has been calculated on y axis and x axis and produces similar plot.

Given the threshold $th = 0.5$, 87.5% of pointed corners can be considered as being homogeneous over the group of labelling people. The overall mean standard deviation on this page reaches 0,15% (1.13 pixels) for the width, and 0.12% (1.28 pixels) for the height. The two bumps, at -40 and 15, are related to the poor segmentation of 13 of the 80 panels. Indeed, instructions have been misunderstood by some people who included text area outside of the panels or missed some panel's parts. Figure D.8 shows the difference between areas labelled as a panel by at least one person and areas labelled as a panel by every participants. However, such mistakes have been manually corrected before publishing the ground truth.

Even though the error criterion has only been estimated on panels, it is reasonable to extend it to balloons and text lines as well. Indeed, the segmentation protocol being quite similar for all features (bounding box as close as possible to the drawing), the observed standard deviation of panel corner positions has no reason to be different from balloons and text lines. The pixel-level balloon and the comic characters visual annotation have been carried out by a single person, the homogeneity is only subject to the regularity of the person over time and is, therefore, difficult to assess quantitatively.

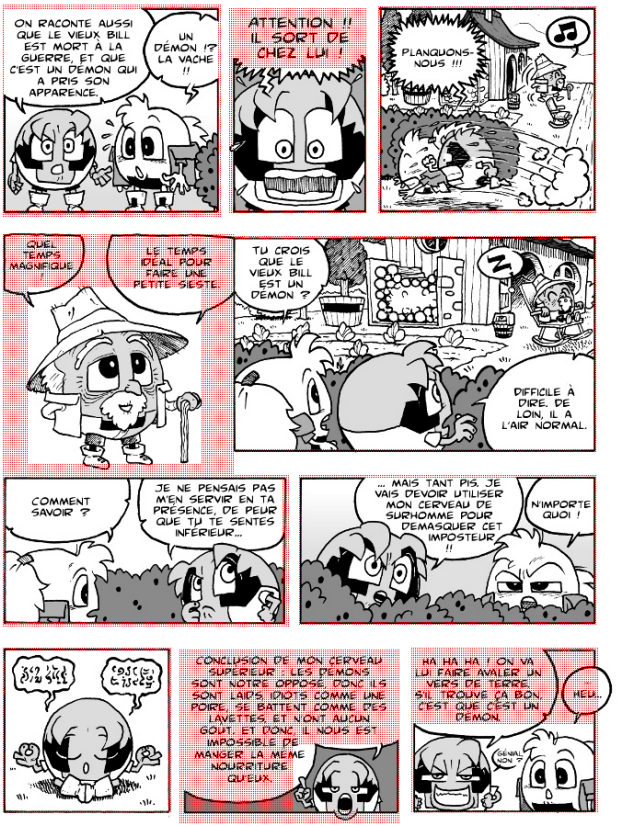


Figure D.8: Image used for error measurement. Red hatched areas are the difference between areas labelled as panels by at least one person and areas labelled by everybody. Image credit: [38].

D.3 Terms of use

We obtain the minimum rights for sharing and publishing image material from the right holders but we had to make sure the user accept it before using the data. In collaboration with the intellectual property department of the University of La Rochelle, we established the following:

In order to use this database, you must firstly agree to the terms. You may not use the database if you don't accept the terms. The use of this database is limited to scien-

tific and non-commercial purpose only, in the computer science domain. For instance, you are allowed to split the images, through the use of segmentation algorithms. You can also use pieces of this database to illustrate your research in publications and presentations. Any other use case must be validated by our service. If you do agree to be bound by all of these Terms of Use, please fill and email the request form and then use the login and password provided to download the selected version below.

The concerned request form requires the identity, affiliation, address and intended use of the person who wishes to use data.

Appendix E

List of Publications

This dissertation has led to the following communications:

Journal Papers

- Christophe Rigaud, Clément Guérin, Dimosthenis Karatzas, Jean-Christophe Burie and Jean-Marc Ogier. "Knowledge-driven understanding of images in comic books". International Journal on Document Analysis and Recognition (IJDAR), 2015 (accepted with minor reviews).

Book series

- Christophe Rigaud, Dimosthenis Karatzas, Jean-Christophe Burie and Jean-Marc Ogier. "Adaptive contour classification of comics speech balloons". In Graphic Recognition. New Trends and Challenges. Lecture Notes in Computer Science (LNCS), Vol. 8746, 2014.

- Christophe Rigaud, Norbert Tsopze, Jean-Christophe Burie and Jean-Marc Ogier. "Robust frame and text extraction from comic books". In Graphic Recognition. New Trends and Challenges. Lecture Notes in Computer Science (LNCS), Vol. 7423, pp. 129-138, 2013.

Conferences

- Clément Guérin, Christophe Rigaud, Karell Bertet, Jean-Christophe Burie, Arnaud Revel and Jean-Marc Ogier. "Réduction de l'espace de recherche pour les personnages de bandes dessinées". In the Proceedings of the 19ème congrès national sur la Reconnaissance de Formes et l'Intelligence Artificielle (RFIA), Rouen, France, July, 2014.

- Christophe Rigaud, Dimosthenis Karatzas, Jean-Christophe Burie and Jean-Marc Ogier. "Color descriptor for content-based drawing retrieval". In the Proceedings of the 11th IAPR International Workshop on Document Analysis Systems (DAS), pp. 267-271 , Tours, France, April, 2014.

- Christophe Rigaud, and Clément Guérin. "Localisation contextuelle des personnages de bandes dessinées". In the Proceedings of the 13ème Colloque International Francophone sur l'Ecrit et le Document (CIFED), pp. 367–370, Nancy, France, March 2014.

- Clément Guérin, Christophe Rigaud, Antoine Mercier, Farid Ammar-Boudjelal, Karell Bertet, Alain Bouju, Jean-Christophe Burie, Georges Louis, Jean-Marc Ogier and Arnaud Revel. "eBDtheque: a representative database of comics". In the Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 1145-1149, Washington DC, USA, August, 2013.

- Christophe Rigaud, Dimosthenis Karatzas, Joost Van de Weijer, Jean-Christophe Burie and Jean-Marc Ogier. "An active contour model for speech balloon detection in comics". In the Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 1240-1244, Washington DC, USA, August, 2013.

- Christophe Rigaud, Dimosthenis Karatzas, Jean-Christophe Burie and Jean-Marc Ogier. "Speech balloon contour classification in comics". In the Proceedings of the 10th IAPR International Workshop on Graphics RECognition (GREC), pp. 23-25, Bethlehem, PA, USA, August, 2013.

- Hoang Nam Ho, Christophe Rigaud, Jean-Christophe Burie and Jean-Marc Ogier. "Redundant structure detection in attributed adjacency graphs for character detection in comics books". In the Proceedings of the 10th IAPR International Workshop on Graphics RECognition (GREC), pp. 109-113, Bethlehem, PA, USA, August, 2013.

- Christophe Rigaud, Dimosthenis Karatzas, Joost Van de Weijer, Jean-Christophe Burie and Jean-Marc Ogier. "Automatic Text Localisation in Scanned Comic Books". In the Proceedings of the 8th International Conference on Computer

Vision Theory and Applications (VISAPP), pp. 814-819, Barcelona, Spain, February, 2013.

- Christophe Rigaud, Norbert Tsopze, Jean-Christophe Burie and Jean-Marc Ogier. "Extraction robuste des cases et du texte de bandes dessinées". In the Proceedings of the 10ème Colloque International Francophone sur l'Ecrit et le Document (CIFED), pp. 349-360, Bordeaux, France, March 2012.

Websites

- eBDtheque project: `http://l3i.univ-larochelle.fr/eBDtheque`
- eBDtheque dataset: `http://ebdtheque.univ-lr.fr`
- Christophe Rigaud's homepage: `http://www.christophe-rigaud.com`

Social networks

- ResearchGate: `http://www.researchgate.net/profile/Christophe_Rigaud`
- GitHub: `https://github.com/crigaud/publication`
- Mendeley: `http://www.mendeley.com/profiles/christophe-r/`

Bibliography

[1] Sadegh Abbasi, Farzin Mokhtarian, and Josef Kittler. Curvature scale space image in shape similarity retrieval. *Multimedia Syst.*, 7(6):467–476, November 1999. 15, 16

[2] Rolf Adams and Leanne Bischof. Seeded region growing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(6):641–647, 1994. 125

[3] Hafiz Aziz AHMAD, Shinichi KOYAMA, and Haruo HIBINO. Impacts of manga on indonesian readers self-efficacy and behavior intentions to imitate its visuals. *Bulletin of JSSD*, 59(3), 2012. 18

[4] Tiago Alves, Ana Simões, Rui Figueiredo, Marco Vala, Ana Paiva, and Ruth Aylett. So tell me what happened: Turning agent-based interactive drama into comics. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 3*, AAMAS '08, pages 1269–1272, Richland, SC, 2008. International Foundation for Autonomous Agents and Multiagent Systems. 20

[5] C-N E. Anagnostopoulos, Ioannis E Anagnostopoulos, Ioannis D Psoroulas, Vassilis Loumos, and Eleftherios Kayafas. License plate recognition from still images and video sequences: A survey. *Intelligent Transportation Systems, IEEE Transactions on*, 9(3):377–391, 2008. 17

[6] Kohei Arai and Herman Tolle. Automatic e-comic content adaptation. *International Journal of Ubiquitous Computing (IJUC)*, 1(1):1–11, 2010. 14

[7] Kohei Arai and Herman Tolle. Method for automatic e-comic scene frame extraction for reading comic on mobile devices. In *Seventh International Conference on Information Technology: New Generations*, ITNG '10, pages 370–375, Washington, DC, USA, 2010. IEEE Computer Society. 13, 90, 91, 98

[8] Kohei Arai and Herman Tolle. Method for real time text extraction of digital manga comic. *International Journal of Image Processing (IJIP)*, 4(6):669–676, 2011. 6, 13, 14, 15, 17, 47, 49, 51, 94, 95, 97, 99, 101, 102

[9] M. Back, R. Gold, A. Balsamo, M. Chow, M. Gorbet, S. Harrison, D. MacDonald, and S. Minnerman. Designing innovative reading experiences for a museum exhibition. *Computer*, 34(1):80–87, Jan 2001. 3

[10] Thomas Bader, René Räpple, and Jürgen Beyerer. Fast invariant contour-based classification of hand symbols for hci. In Xiaoyi Jiang and Nicolai Petkov, editors, *Computer Analysis of Images and Patterns*, volume 5702 of *Lecture Notes in Computer Science*, pages 689–696. Springer Berlin Heidelberg, 2009. 15

[11] Regina Bernhaupt and Sandra SchonetSun. Using artificial neural networks as an image segmentation module of an ocr-system: A preliminary study. In *Proceedings of 5th WSES/IEEE World Multiconference*, pages 6751–6756, Island of Crete, Greece, 2001. 17, 18

[12] Siddhartha Bhattacharyya. A brief survey of color image preprocessing and segmentation techniques. *Journal of Pattern Recognition Research*, 1(1):120–129, 2011. 10

[13] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. 127

[14] Thomas Blaschke, Geoffrey J. Hay, Maggi Kelly, Stefan Lang, Peter Hofmann, Elisabeth Addink, Raul Queiroz Feitosa, Freek van der Meer, Harald van der Werff, Frieke van Coillie, and Dirk Tiede. Geographic object-based image analysis – towards a new paradigm. {*ISPRS*} *Journal of Photogrammetry and Remote Sensing*, 87(0):180 – 191, 2014. 19

[15] Miroslaw Bober. Mpeg-7 visual shape descriptors. *IEEE Trans. Circuits Systems*, 2001. 16

[16] Michał Borodo. Multimodality, translation and comics. *Perspectives*, pages 1–20, 2014. 3

[17] DC Brandon. Graphic novels and comics for the visually impaired explored in award-winning paper, 2014. 3

[18] Thomas M Breuel. High performance document layout analysis. In *Proceedings of the Symposium on Document Image Understanding Technology*, pages 209–218, 2003. 11

[19] Christopher Brochtrup. Ocr manga reader for android, 2013. 20

[20] Christopher Brochtrup. Capture2text, 2014. 20

[21] Chrysoline Canivet-Fovez. *Le manga Une synthèse de référence qui éclaire en image l'origine, l'histoire et l'influence de la bande dessinée japonaise*. Eyrolles, 2014. 2

[22] John Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8(6):679–698, Nov 1986. 126

[23] ChungHo Chan, Howard Leung, and Taku Komura. Automatic panel extraction of color comic images. In HoraceH.-S. Ip, OscarC. Au, Howard Leung, Ming-Ting Sun, Wei-Ying Ma, and Shi-Min Hu, editors, *Advances in Multimedia*

Information Processing - PCM 2007, volume 4810 of *Lecture Notes in Computer Science*, pages 775–784. Springer Berlin Heidelberg, 2007. 13

[24] S.C.S. Cheung, City University of Hong Kong. Run Run Shaw Library, and City University of Hong Kong. Department of Computer Science. *Face Detection and Face Recognition of Human-like Characters in Comics*. Outstanding academic papers by students. Run Run Shaw Library, City University of Hong Kong, 2008. 18

[25] G. Chowdhury. *Introduction to Modern Information Retrieval, Third Edition*. Facet Publishing, 3rd edition, 2010. 61

[26] Evans Christophe and Gaudet Francoise. La lecture de bandes dessinées en france [ce-2012-2]. Technical report, Ministère de la Culture et de la Communication, Paris, France, 2012. 2

[27] Marx Christy. *Writing for Animation, Comics, and Games*, chapter 5, pages 79–132. Focal Press, Taylor and Francis Group, Waltham, Massachusetts, Middlesex, US, 1 edition, 2007. 2

[28] Wei-Ta Chu and Chia-Hsiang Yu. Optimized speech balloon placement for automatic comics generation. In *Proceedings of the 3rd ACM International Workshop on Interactive Multimedia on Mobile and Portable Devices*, IMMPD '13, pages 1–6, New York, NY, USA, 2013. ACM. 20

[29] Antonio Clavelli and Dimosthenis Karatzas. Text segmentation in colour posters from the spanish civil war era. In *Proceedings of International Conference on Document Analysis and Recognition*, ICDAR '09, pages 181–185, Washington, DC, USA, 2009. IEEE Computer Society. 5, 17, 25

[30] Laurent D Cohen. On active contour models and balloons. *CVGIP: Image Understanding*, 53(2):211–218, March 1991. 15

[31] Neil Cohn. The limits of time and transitions: Challenges to theories of sequential image comprehension. *Studies in Comics*, 1(1):127–147, 2010. 18

[32] Neil Cohn. Navigating comics: an empirical and theoretical approach to strategies of reading comic page layouts. *Frontiers in psychology*, 4(April):186, 2013. 14

[33] International Committee. *Introduction to publishing in Japan*. Japan Book Publishers Association, Tokyo, Japan, 2014. 2

[34] Robert S Cooperman. System for document layout analysis, 1998. US Patent 5,784,487. 11

[35] Rikke Platz Cortsen. *Comics as Assemblage: How spatio-temporality in comics is constructed*. PhD thesis, University of Copenhagen, 2012. 3

[36] Daniel Cremers, Florian Tischhäuser, Joachim Weickert, and Christoph Schnörr. Diffusion snakes: introducing statistical shape knowledge into the mumford-shah functional. *International Journal of Computer Vision*, 50:295–313, 2002. 30

[37] Cyb. *Cosmozone*. Studio Cyborga, Goven, France, 2008. 59, 60, 64

[38] Cyb. *Bubblegôm Gôm*. Studio Cyborga, Goven, France, 2009. 14, 23, 26, 27, 29, 31, 33, 49, 149

[39] Cyb. *Le Magicien Lose*. Studio Cyborga, Goven, France, 2011. 68, 109

[40] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In Cordelia Schmid, Stefano Soatto, and Carlo Tomasi, editors, *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 886–893, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, June 2005. 18

[41] Danica Davidson. Manga grows in the heart of europe, 2012. 2

[42] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18, 2000. 48

[43] Adrien Delaye and Cheng-Lin Liu. Multi-class segmentation of free-form online documents with tree conditional random fields. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–17, 2014. 5

[44] Michael B Dillencourt, Hanan Samet, and Markku Tamminen. A general approach to connected-component labeling for arbitrary image representations. *Journal of the ACM (JACM)*, 39(2):253–280, 1992. 129

[45] Philippe Dosch, Karl Tombre, Christian Ah-Soon, and Gérald Masini. A complete system for the analysis of architectural drawings. *International Journal on Document Analysis and Recognition*, 3(2):102–116, 2000. 19

[46] B. Duc. *L'art de la B.D.: Du scénario à la réalisation graphique, tout sur la création des bandes dessinées*. Editions Glénat, 1997. 1, 5, 76

[47] Richard O. Duda and Peter E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM*, 15:11–15, January 1972. 13

[48] Øivind Due Trier, Anil K Jain, and Torfinn Taxt. Feature extraction methods for character recognition-a survey. *Pattern recognition*, 29(4):641–662, 1996. 10

[49] Will Eisner. *Comics and Sequential Art*. W. W. Norton and Company, 1985. 2

[50] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2963 –2970, june 2010. 5, 17

[51] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 85, 139

[52] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 88

[53] S. Fidler, Jian Yao, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 0:702–709, 2012. 19

[54] Lloyd A. Fletcher and Rangachar Kasturi. A robust algorithm for text string separation from mixed text/graphics images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(6):910–918, 1988. 17

[55] Joyce Goggin. *The Rise and Reason of Comics and Graphic Literature: Critical Essays on the Form*, volume 1, chapter 4, pages 56–74. McFarland, Book News, Inc., Portland, OR, 1 edition, 2010. 14

[56] Lluis Gomez and Dimosthenis Karatzas. Multi-script text extraction from natural scenes. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 467–471. IEEE, 2013. 16

[57] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2nd edition, 1992. 126

[58] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. 85

[59] A Grigoriu, JP Vonwiller, and RW King. An automatic intonation tone contour labelling and classification algorithm. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, volume 2, pages II–181. IEEE, 1994. 15

[60] Laurence Grove. Bande dessinée studies. *French Studies*, 68(1):78–87, 2014. 2

[61] Clément Guérin. Ontologies and spatial relations applied to comic books reading. In *PhD Symposium of Knowledge Engineering and Knowledge Management (EKAW)*, Galway, Ireland, 2012. 14, 19, 68

[62] Clément Guérin. *Enrichissement sémantique d'images de planches de bandes déssinées numérisées*. PhD thesis, Université de La Rochelle, Technoforum - 23, avenue Albert Einstein BP 33060 - 17031 La Rochelle - France, 2014. 68, 70, 139, 144

[63] Clément Guérin, Christophe Rigaud, Antoine Mercier, and al. ebdtheque: a representative database of comics. In *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*, Washington DC, 2013. 86, 89, 90, 94, 97, 98, 101, 107, 115, 116

[64] Volker Haarslev, Kay Hidde, Ralf Möller, and Michael Wessel. The RacerPro knowledge representation and reasoning system. *Semantic Web*, 3(3):267–277, 2012. 74

[65] Eunjung Han, Kirak Kim, HwangKyu Yang, and Keechul Jung. Frame segmentation used mlp-based x-y recursive for mobile cartoon content. In *Proceedings of the 12th international conference on Human-computer interaction: intelligent multimodal interaction environments*, HCI'07, pages 872–881, Berlin, Heidelberg, 2007. Springer-Verlag. 13

[66] Alice Hermann, Sébastien Ferré, and Mireille Ducassé. Guided semantic annotation of comic panels with sewelis. In *EKAW*, volume 7603 of *Lecture Notes in Computer Science*, pages 430–433. Springer, 2012. 19

[67] Anh Khoi Ngo Ho, Jean-Christophe Burie, and Jean-Marc Ogier. Panel and speech balloon extraction from comic books. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 424–428. Ieee, 2012. xi, 6, 13, 14, 47, 90, 91, 92, 94, 98, 99, 102

[68] Hoang Nam Ho, Christophe Rigaud, Jean-Christophe Burie, and Jean-Marc Ogier. Redundant structure detection in attributed adjacency graphs for character detection in comics books. In *Proceedings of the 10th IAPR International Workshop on Graphics Recognition (GREC)*, Bethlehem, PA, USA, 2013. 18

[69] Steven L Horowitz and Theodosios Pavlidis. Picture segmentation by a tree traversal algorithm. *Journal of the ACM (JACM)*, 23(2):368–388, 1976. 126

[70] B. Hu, S. Dasmahapatra, P. Lewis, and N. Shadbolt. Ontology-based medical image annotation with description logics. In *Tools with Artificial Intelligence, 2003. Proceedings of the 15th Int. Conf. on Tools with Artificial Intelligence (ICTAI'03)*, 2003. 19

[71] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on*, 8(2):179–187, 1962. 16

[72] Céline Hudelot, Jamal Atif, and Isabelle Bloch. Fuzzy spatial relation ontology for image interpretation. *Fuzzy Sets and Systems*, 159(15):1929 – 1951, 2008. From Knowledge Representation to Information Processing and Management Selected papers from the French Fuzzy Days (LFA 2006). 19

[73] Yusuke In, Takashi Oie, Masakazu Higuchi, Shuji Kawasaki, Atushi Koike, and Hitomi Murakami. Fast frame decomposition and sorting by contour tracing for mobile phone comic images. *Internatinal journal of systems applications, engineering and development*, 5(2):216–223, 2011. 13

[74] Eri Inoue. *Kyoumen gensoukyoku*. Kyoumen gensoukyoku, Osaka, Japan, 2008. 14, 26, 27, 49

[75] Motoi Iwata, Atsushi Ito, and Koichi Kise. A study to achieve manga character retrieval method for manga images. In *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*, pages 309–313, April 2014. 18

[76] Raulet Jérémy and Boyer Vincent. Comics reading: An automatic script generation. In *Proceedings of the 21st International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, pages 88–96, 2013. 3

[77] Olivier Jolivet. *BostonPolice*. Clair de Lune, Allauch, France, 2010. 59, 142

[78] Llados Josep. Recognition of textual and graphical patterns. Second International Document Image Processing Summer School (IDIPS 2014), 2014. 11

[79] Keechul Jung, Kwang In Kim, and Anil K. Jain. Text information extraction in images and video: a survey. *Pattern Recognition*, 37(5):977 – 997, 2004. 16

[80] D. Karatzas and A. Antonacopoulos. Colour text segmentation in web images based on human perception. *Image and Vision Computing*, 25(5):564 – 577, 2007. 17

[81] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. Gomez i Bigorda, S. Robles Mestre, J. Mas, D. Fernandez Mota, J. Almazan Almazan, and L.-P. de las Heras. Icdar 2013 robust reading competition. In *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*, pages 1484–1493, Aug 2013. 16

[82] Thotreingam Kasar and Angarai G Ramakrishnan. Multi-script and multi-oriented text localization from scene images. In *Camera-Based Document Analysis and Recognition*, pages 1–14. Springer, 2012. 16

[83] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988. 15, 30, 126

[84] D. Kelkar and S. Gupta. Improved quadtree method for split merge image segmentation. In *Emerging Trends in Engineering and Technology, 2008. ICETET '08. First International Conference on*, pages 44–47, July 2008. 126

[85] Eamonn Keogh, Li Wei, Xiaopeng Xi, Sang hee Lee, and Michail Vlachos. Lb keogh supports exact indexing of shapes under rotation invariance with arbitrary representations and distance measures. In *IN VLDB, 2006*, pages 882–893, 2006. 16

[86] Fahad Shahbaz Khan, Muhammad Anwer Rao, Joost van de Weijer, Andrew D. Bagdanov, Maria Vanrell, and Antonio Lopez. Color attributes for object detection. In *Twenty-Fifth IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, 2012. 18

[87] Wonjun Kim and Changick Kim. A new approach for overlay text detection and extraction from complex video scene. *Image Processing, IEEE Transactions on*, 18(2):401 –411, feb. 2009. 17

[88] Johannes Kopf and Dani Lischinski. Digital reconstruction of halftoned color comics. *ACM Trans. Graph.*, 31(6):140:1–140:10, November 2012. 59, 113

[89] Gerald Kühne, Stephan Richter, and Markus Beier. Motion-based segmentation and contour-based classification of video objects. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 41–50. ACM, 2001. 15

[90] David Kurlander, Tim Skelly, and David Salesin. Comic chat. *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques - SIGGRAPH '96*, 96:225–236, 1996. 20

[91] Jean-Marc Lainé and Sylvain Delzant. *Le lettrage des bulles*. Eyrolles, 2010. 5, 76

[92] Bart Lamiroy, Daniel Lopresti, Hank Korth, and Jeff Heflin. How Carefully Designed Open Resource Sharing Can Help and Expand Document Analysis Research. In Christian Viard-Gaudin Gady Agam, editor, *Document Recognition and Retrieval XVIII - DRR 2011*, volume 7874, San Francisco, United States, January 2011. SPIE, SPIE. 85

[93] Bart Lamiroy and Jean-Marc Ogier. Analysis and interpretation of graphical documents. In David Doermann and Karl Tombre, editors, *Handbook of Document Image Processing and Recognition*. Springer, 2014. 9, 19

[94] Lamisseb. *Les noeils Tome 1.* Bac@BD, Valence, France, 2011. 59, 60, 62, 63

[95] Lamisseb. *Et Pis Taf !* Bac@BD, Valence, France, 2012. 34

[96] Bertrand Leroy, IsabelleL. Herlin, and LaurentD. Cohen. Multi-resolution algorithms for active contour models. In Marie-Odile Berger and Deriche et al., editors, *ICAOS'96*, volume 219 of *Lecture Notes in Control and Information Sciences*, pages 58–65. Springer Berlin Heidelberg, 1996. 15

[97] Wing Ho Leung and Tsuhan Chen. Trademark retrieval using contour-skeleton stroke classification. In *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, volume 2, pages 517–520. IEEE, 2002. 15

[98] Congcong Li, Adarsh Kowdle, Ashutosh Saxena, and Tsuhan Chen. Toward holistic scene understanding: Feedback enabled cascaded classification models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(7):1394–1408, 2012. 19

[99] Luyuan Li, Yongtao Wang, Zhi Tang, and Liangcai Gao. Automatic comic page segmentation based on polygon detection. *Multimedia Tools Applications*, 69(1):171–197, 2014. 13

[100] Luyuan Li, Yongtao Wang, Zhi Tang, and Dong Liu. Comic image understanding based on polygon detection. *Proc. SPIE*, 8658:86580B–86580B–11, 2013. 14, 25

[101] Luyuan Li, Yongtao Wang, Zhi Tang, Xiaoqing Lu, and Liangcai Gao. Unsupervised speech text localization in comic images. In *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*, pages 1190–1194, Aug 2013. 17, 46, 48

[102] Bruce Lidl. Surveying digital comics after amazon-comixology, 2014. 4

[103] Dong Liu, Yongtao Wang, Zhi Tang, Luyuan Li, and Liangcai Gao. Automatic comic page image understanding based on edge segment analysis. In *Proc. SPIE*, volume 9021, pages 90210J–90210J–12, 2013. 14

[104] H-C Liu and Mandyam D. Srinath. Partial shape classification using contour matching in distance transformation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(11):1072–1079, 1990. 15

[105] Martin Lopatka and Wiger Van Houten. Science and Justice Automated shape annotation for illicit tablet preparations: a contour angle based classification from digital images. *Science & Justice*, 53(1):60–66, 2013. 16

[106] Ricardo Lopes, Tiago Cardoso, Nelson Silva, and ManuelJ. Fonseca. Calligraphic shortcuts for comics creation. In Andreas Butz, Brian Fisher, Marc Christie, Antonio Krüger, Patrick Olivier, and Roberto Therón, editors, *Smart Graphics*, volume 5531 of *Lecture Notes in Computer Science*, pages 223–232. Springer Berlin Heidelberg, 2009. 20

[107] Gérald Lubbin. *Les bulles du labo.* Doc en Stock, Paris, France, 2012. 45, 59

[108] L Luccheseyz and SK Mitray. Color image segmentation: A state-of-the-art survey. *Proceedings of the Indian National Science Academy (INSA-A)*, 67(2):207–221, 2001. 10

[109] Boudissa Magali. *La bande dessinée entre la page et l'écran : étude critique des enjeux théoriques liés au renouvellement du langage bédéique sous influence numérique.* PhD thesis, University Paris 8 Vincennes Saint-Denis, Paris, France, 2011. 2

[110] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936. 48

[111] Song Mao, Azriel Rosenfeld, and Tapas Kanungo. Document structure analysis algorithms: a literature survey. In Tapas Kanungo, Elisa H. Barney Smith, Jianying Hu, and Paul B. Kantor, editors, *DRR*, volume 5010 of *SPIE Proceedings*, pages 197–207. SPIE, 2003. 19

[112] Christy Marx. *Writing for Animation, Comics, and Games.* Focal Press, 2006. 14

[113] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767, 2004. 18, 48

[114] Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa. Interactive Manga retargeting. In *ACM SIGGRAPH 2011 Posters on - SIGGRAPH '11*, page 1, New York, New York, USA, 2011. ACM Press. 13, 14, 17

[115] Ted McCall and Charles Snelgrove. *Robin Hood And Company.* Anglo-American, Canada, 1946. 14, 27, 49

[116] Scott McCloud. *Understanding comics: The Invisible Art.* Kitchen Sink Press, 1993. 1, 2, 86

[117] Scott McCloud. *Making Comics: Storytelling Secrets of Comics, Manga and Graphic Novels.* HarperCollins, 2006. 21

[118] Jason McIntosh. ComicsML, 2011. 144

[119] Stuart Medley. Discerning pictures: How we look at and understand images in comics. *Studies in Comics*, 1(1):53–70, 2010. 18

[120] Andrew Mehnert and Paul Jackway. An improved seeded region growing algorithm. *Pattern Recognition Letters*, 18(10):1065–1071, 1997. 125

[121] Quan Meng and Yonghong Song. Text detection in natural scenes with salient region. In *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*, pages 384 –388, march 2012. 17

[122] Aaron Meskin and Roy T. Cook. *The Art of Comics: A Philosophical Approach - Ch 2 The Ontology of Comics*, volume 1, pages 218–285. John Wiley and Sons, New York, USA, 2011. 19, 68

[123] Vasileios Mezaris, Ioannis Kompatsiaris, and Michael G. Strintzis. An ontology approach to object-based image retrieval. In *In Proc. IEEE Int. Conf. on Image Processing (ICIP03*, pages 511–514, 2003. 19

[124] Midam. *Kid Paddle*. Dupuis, Marcinelle, Belgium, 2011. 28, 52, 60

[125] John Jackson Miller. Yearly rankings for comic book sales, 2014. 3

[126] G.S. Millidge. *Comic Book Design*. Watson-Guptill Publications, 2009. 16

[127] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997. 57

[128] Farzin Mokhtarian and Sadegh Abbasi. Shape similarity retrieval under affine transforms. *Pattern Recognition*, 35(1):31 – 41, 2002. 16

[129] Marion Montaigne. *Tu mourras moins bête - La science, c'est pas du cinéma !* Ankama, Roubaix, France, 2012. 34

[130] Ramakrishnan Mukundan and KR Ramakrishnan. *Moment functions in image analysis: theory and applications*, volume 100. World Scientific, 1998. 15

[131] George Nagy. Twenty years of document image analysis in pami. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):38–62, 2000. 17

[132] L. Neumann and J. Matas. Real-time scene text localization and recognition. *Computer Vision and Pattern Recognition*, pages 1485–1490, 2012. 5, 17

[133] Nhu Van Nguyen, Jean-Marc Ogier, and Franck Charneau. Bag of subjects: Lecture videos multimodal indexing. In *Proceedings of the 2013 ACM Symposium on Document Engineering*, DocEng '13, pages 225–226, New York, NY, USA, 2013. ACM. 17

[134] Alain Saint Ogan. *Zig, Puce et Alfred*. Hachette, Paris, France, 1952. 49

[135] J.M. Ogier, R. Mullot, J. Labiche, and Y. Lecourtier. Semantic coherency: the basis of an image interpretation device-application to the cadastral map interpretation. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 30(2):322–338, Apr 2000. 19

[136] Daniel M. Oliveira and Rafael D. Lins. Generalizing tableau to any color of teaching boards. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, ICPR '10, pages 2411–2414, Washington, DC, USA, 2010. IEEE Computer Society. 5, 17

[137] N. Otsu. A threshold selection method from gray level histograms. *IEEE Trans. Systems, Man and Cybernetics*, 9:62–66, March 1979. minimize inter class variance. 51, 127

[138] Nikhil R Pal and Sankar K Pal. A review on image segmentation techniques. *Pattern recognition*, 26(9):1277–1294, 1993. 10

[139] Xufang Pang, Ying Cao, Rynson W.H. Lau, and Antoni B. Chan. A robust panel extraction method for manga. In *Proceedings of the ACM International Conference on Multimedia*, MM '14, pages 1125–1128, New York, NY, USA, 2014. ACM. 13

[140] B Pasternak. Processing imprecise and structural distorted line drawings by an adaptable drawing interpretation kernel. In *Proc. of IAPR Workshop on Document Analysis Systems*, pages 349–366, Kaiserslautern, Germany, 1994. 19

[141] B. Pasternak. The role of taxonomy in drawing interpretation. In *Proc. of the Third International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 799–802 vol.2, Aug 1995. 19

[142] Jay Scott Pike. *Boy Love Girls*. Lev Gleason, New York City, US, 1953. 77

[143] Plasq. Comic life 3, 2014. 20

[144] J Ponce and D Forsyth. *Computer vision: a modern approach*. Prentice Hall (2nd edition), 2012. 127

[145] Christophe Ponsard. Enhancing the accessibility for all of digital comic books. *e-Minds*, 1(5), 2009. 3, 14

[146] Christophe Ponsard, Ravi Ramdoyal, and Daniel Dziamski. An ocr-enabled digital comic books viewer. In *Computers Helping People with Special Needs*, pages 471–478. Springer, 2012. 13, 17

[147] Vicky Portail-Kernel and Cédric Kernel. *Prunelle, la fille du cyclope*. Ankama, Roubaix, France, 2010. 18, 60

[148] Pratt and William K. *Digital image processing (2nd ed.)*. John Wiley & Sons, Inc., New York, NY, USA, 1991. 22

[149] William K. Pratt. *Digital Image Processing: PIKS Inside*. John Wiley & Sons, Inc., New York, NY, USA, 4th edition, 2001. 125

[150] Gilles Ratier. 2013 : l'année de la décélération, 2013. 2

[151] J. Raulet and V. Boyer. A sketch-based interface to script comics reading. In *SIGGRAPH Asia 2011 Sketches*, SA '11, pages 3:1–3:2, New York, NY, USA, 2011. ACM. 20

[152] Stephan Richter, Gerald Kühne, and Oliver Schuster. Contour-based classification of video objects. In *Proceedings of SPIE*, volume 4315, page 608, 2001. 15

[153] Paul Rissen. A Comics Ontology, 2012. 144

[154] Gibbons Robin Varnum, Christina T. *The Language of Comics: Word and Image*. Studies in Popular Culture. University Press of Mississippi, 2007. 14, 35

[155] P.P. Roy, J. Llados, and U. Pal. Text/graphics separation in color maps. In *Computing: Theory and Applications, 2007. ICCTA '07. International Conference on*, pages 545–551, March 2007. 17

[156] P.P. Roy, U. Pal, J. Llados, and M. Delalandre. Multi-oriented and multi-sized touching character segmentation using dynamic programming. In *10th International Conference on Document Analysis and Recognition (ICDAR)*, pages 11–15, July 2009. 26

[157] Antti Salovaara. Appropriation of a mms-based comic creator: From system functionalities to resources for action. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 1117–1126, New York, NY, USA, 2007. ACM. 20

[158] Sohail Sarwar, Zia Ul Qayyum, and Saqib Majeed. Ontology based image retrieval framework using qualitative semantic image descriptions. *Procedia Computer Science*, 22(0):285 – 294, 2013. 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - {KES2013}. 19

[159] Eugenio Di Sciascio, Francesco M. Donini, and Marina Mongiello. Structured knowledge representation for image retrieval. *CoRR*, abs/1109.1498, 2011. 19, 74

[160] Mehmet Sezgin et al. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging*, 13(1):146–168, 2004. 127

[161] A Shahab, F. Shafait, and A Dengel. Icdar 2011 robust reading competition challenge 2: Reading text in scene images. In *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*, pages 1491–1496, Sept 2011. 16

[162] Ariel Shamir, Michael Rubinstein, and Tomer Levinboim. Generating comics from 3d interactive computer graphics. *IEEE Comput. Graph. Appl.*, 26(3):53–61, May 2006. 20

[163] P. Shivakumara, T.Q. Phan, and Chew Lim Tan. A robust wavelet transform based technique for video text detection. In *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*, pages 1285 –1289, july 2009. 17

[164] Shueisha. Vomic, 2014. 20

[165] Siddharth Singh, Adrian David Cheok, Guo Loong Ng, and Farzam Farbiz. 3d augmented reality comic book and notes for children using mobile phones. In *Proceedings of the 2004 Conference on Interaction Design and Children: Building a Community*, IDC '04, pages 149–150, New York, NY, USA, 2004. ACM. 3

[166] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. Pellet : A Practical OWL-DL Reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):51–53, 2007. 75

[167] A.F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 321–330. ACM, 2006. 85

[168] Martin Stommel, Lena I Merhej, and Marion G Müller. Segmentation-free detection of comic panels. In *Computer Vision and Graphics*, pages 633–640. Springer, 2012. 13

[169] Chung-Yuan Su, Ray-I Chang, and Jen-Chang Liu. Recognizing text elements for svg comic compression and its novel applications. In *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*, pages 1329–1333, Washington, DC, USA, 2011. IEEE Computer Society. 17

[170] Yasuyuki Sumi, Ryuuki Sakamoto, Keiko Nakao, and Kenji Mase. Comicdiary: Representing individual experiences in a comics style. In *Ubicomp*, pages 16–32, 2002. 20

[171] Kang B Sun and Boaz J Super. Classification of contour shapes using class segment sets. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 727–733. IEEE, 2005. 15

[172] Weihan Sun and K. Kise. Detecting printed and handwritten partial copies of line drawings embedded in complex backgrounds. In *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*, pages 341–345, July 2009. 18

[173] Weihan Sun and Koichi Kise. Similar partial copy detection of line drawings using a cascade classifier and feature matching. In Hiroshi Sako, Katrin Franke, and Shuji Saitoh, editors, *ICWF*, volume 6540 of *Lecture Notes in Computer Science*, pages 126–137. Springer, 2010. 18

[174] Weihan Sun and Koichi Kise. Similar Manga Retrieval Using Visual Vocabulary Based on Regions of Interest. In *2011 International Conference on Document Analysis and Recognition*, pages 1075–1079. Ieee, 2011. 14

[175] Weihan Sun and Koichi Kise. Detection of exact and similar partial copies for copyright protection of manga. *International Journal on Document Analysis and Recognition (IJDAR)*, 16(4):331–349, 2013. 18

[176] Weihan Sun, Koichi Kise, Jean-Christophe Burie, and Jean-Marc Ogier. Specific comic character detection using local feature matching. In *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*, Washington, USA, 2013. 19

[177] M. Sundaresan and S. Ranjini. Text extraction from digital english comic image using two blobs extraction method. In *Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on*, pages 449–452, March 2012. 17

[178] Satoshi Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30(1):32–46, 1985. 129

[179] Saira Syed. Comic giants battle for readers, 2012. 2

[180] Richard Szeliski. *Computer Vision: Algorithms and Applications*, chapter 8, pages 201–213. Springer-Verlag New York, Inc., New York, NY, USA, 1 edition, 2010. 129

[181] Richard Szeliski. *HORS-THÈME/OPEN TOPIC*, chapter 2, pages 80–89. University of Alberta, Edmonton, Canada, 6 edition, 2013. 3

[182] Anh-Phuong TA. *Inexact Graph matching techniques: Application to Object detection and Human action recognition.* PhD thesis, INSA de Lyon, 20, rue Albert Einstein, 69621 Villeurbanne Cedex, France, 2010. 19

[183] Kohei Takayama, Henry Johan, and Tomoyuki Nishita. Face detection and face recognition of cartoon characters using feature extraction. In *Image Electronics and Visual Computing Workshop (IEVC'12)*, Kuching, Malaysia, 2012. 18

[184] Takamasa Tanaka, Kenji Shoji, Fubito Toyama, and Juichi Miyamichi. Layout analysis of tree-structured scene frames in comic images. In *IJCAI'07*, pages 2885–2890, 2007. 13

[185] Yuan Y Tang, Seong-Whan Lee, and Ching Y Suen. Automatic document processing: a survey. *Pattern recognition*, 29(12):1931–1952, 1996. 10

[186] Quang Minh Tieng and WW Boles. Recognition of 2d object contours using the wavelet transform zero-crossing representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(8):910–916, 1997. 15

[187] Hiroaki Tobita. Comic engine: Interactive system for creating and browsing comic books with attention cuing. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '10, pages 281–288, New York, NY, USA, 2010. ACM. 19

[188] Karl Tombre, Salvatore Tabbone, Loïc Pélissier, Bart Lamiroy, and Philippe Dosch. Text/graphics separation revisited. In *Document Analysis Systems V*, pages 200–211. Springer, 2002. 17

[189] Chia-Jung Tsai, Chih-Yuan Yao, Pei-ying Chiang, Yu-Chi Lai, Ming-Te Chi, Hung-Kuo Chu, Yu-Shiang Wong, and Yu-Shuen Wang. Adaptive manga relayout on mobile device. In *ACM SIGGRAPH 2013 Posters*, SIGGRAPH '13, pages 57:1–57:1, New York, NY, USA, 2013. ACM. 13, 14

[190] Shingo Uchihashi, Jonathan Foote, Andreas Girgensohn, and John Boreczky. Video manga: Generating semantically meaningful video summaries. In *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, MULTIMEDIA '99, pages 383–392, New York, NY, USA, 1999. ACM. 20

[191] Szilárd Vajda, Leonard Rothacker, and GernotA. Fink. A method for camera-based interactive whiteboard reading. In Masakazu Iwamura and Faisal Shafait, editors, *Camera-Based Document Analysis and Recognition*, volume 7139 of *Lecture Notes in Computer Science*, pages 112–125. Springer Berlin Heidelberg, 2012. 17

[192] Remco C. Veltkamp and Mirela Tanase. Content-based image retrieval systems: A survey. Technical report, Department of Information and Computing Sciences, Utrecht University, 2000. 16

[193] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004. 18

[194] John A Walsh. Comic Book Markup Language : An Introduction and Rationale. *Digital Humanities Quarterly (DHQ)*, 6(1):1–50, 2012. 144

[195] Kai Wang and Serge Belongie. Word spotting in the wild. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, volume 6311 of *Lecture Notes in Computer Science*, pages 591–604. Springer Berlin / Heidelberg, 2010. 17

[196] Meng Wang, Richang Hong, Xiao-Tong Yuan, Shuicheng Yan, and Tat-Seng Chua. Movie2comics: Towards a lively video content presentation. *IEEE Transactions on Multimedia*, 14(3-2):858–870, 2012. 19

[197] Z. Wang, Z. Chi, and D. Feng. Shape based leaf image retrieval. *Vision, Image and Signal Processing, IEEE Proceedings*, 150(1):34–43, 2003. 16

[198] Coulton Waugh. *The Comics*. Univ. Press of Mississippi, 1947. 2

[199] J.J. Weinman, E. Learned-Miller, and A.R. Hanson. Scene text recognition using similarity and a lexicon with sparse belief propagation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(10):1733 –1746, 2009. 5, 17

[200] C Xu and J L Prince. Snakes, shapes, and gradient vector flow. *IEEE Transactions on Image Processing*, 7(3):359–369, 1998. 15, 30, 102, 126

[201] Li Xu, Cewu Lu, Yi Xu, and Jiaya Jia. Image smoothing via l0 gradient minimization. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, SA '11, pages 174:1–174:12, New York, NY, USA, 2011. ACM. 59, 113

[202] Pengfei Xu, Hongbo Fu, Oscar Kin-Chung Au, and Chiew-Lan Tai. Lazy selection: A scribble-based tool for smart shape elements selection. *ACM Trans. Graph.*, 31(6):142:1–142:9, November 2012. 60

[203] Masashi Yamada, Rahmat Budiarto, Mamoru Endo, and Shinya Miyazaki. Comic image decomposition for reading comics on cellular phones. *IEICE Transactions*, 87-D(6):1370–1376, 2004. 17

[204] B. Yao, X. Yang, and S.C. Zhu. Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 169–183. Springer, 2007. 139

[205] Peng Ye and D. Doermann. Document image quality assessment: A brief survey. In *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*, pages 723–727, Aug 2013. 12

[206] Sungsoo Yoon, Gyeonghwan Kim, Yeongwoo Choi, and Yillbyung Lee. New paradigm for segmentation and recognition of handwritten numeral string. In *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, pages 205–209, 2001. 46

[207] C. T. Zahn and R. Z. Roskies. Fourier Descriptors for Plane Closed Curves. *Transactions on Computers, IEEE*, c-21(3):269–281, March 1972. 15

[208] D. Zhang and G. Lu. Review of shape representation and description techniques. *PR*, 37(1), 2004. 16

[209] S. Zinger, C. Millet, B. Mathieu, G. Grefenstette, P. Hède, and P. a. Moëllic. Extracting an ontology of portrayable objects from wordnet. In *MUSCLE / ImageCLEF workshop on Image and Video retrieval evaluation*, pages 17–23, 2005. 71

Segmentation et indexation d'objets complexes dans les images de bandes dessinées

Résumé :

Dans ce manuscrit de thèse, nous détaillons et illustrons les différents défis scientifiques liés à l'analyse automatique d'images de bandes dessinées, de manière à donner au lecteur tous les éléments concernant les dernières avancées scientifiques en la matière ainsi que les verrous scientifiques actuels.

Nous proposons trois approches pour l'analyse d'image de bandes dessinées. La première approche est dite "séquentielle" car le contenu de l'image est décrit progressivement et de manière intuitive. Dans cette approche, les extractions se succèdent, en commençant par les plus simples comme que les cases, le texte et les bulles qui servent ensuite à guider l'extraction d'éléments plus complexes tels que la queue des bulles et les personnages au sein des cases. La seconde approche propose des extractions indépendantes les unes des autres de manière à éviter la propagation d'erreur due aux traitements successifs. D'autres éléments tels que la classification du type de bulle et la reconnaissance de texte y sont aussi abordés. La troisième approche introduit un système fondé sur une base de connaissance *à priori* du contenu des images de bandes dessinées. Ce système permet de construire une description sémantique de l'image, dirigée par les modèles de connaissances. Il combine les avantages des deux approches précédentes et permet une description sémantique de haut niveau pouvant inclure des informations telles que l'ordre de lecture, la sémantique des bulles, les relations entre les bulles et leurs locuteurs ainsi que les interaction entre les personnages.

Mots clés : traitement d'images, reconnaissance de formes, analyse de documents, compréhension de bandes dessinées.

Segmentation and indexation of complex objects in comic book images

Summary:

In this thesis, we review, highlight and illustrate the challenges related to comic book image analysis in order to give to the reader a good overview about the last research progress in this field and the current issues.

We propose three different approaches for comic book image analysis that are composed by several processing. The first approach is called "sequential" because the image content is described in an intuitive way, from simple to complex elements using previously extracted elements to guide further processing. Simple elements such as panel text and balloon are extracted first, followed by the balloon tail and then the comic character position in the panel. The second approach addresses independent information extraction to recover the main drawback of the first approach, the error propagation. This second method is called "independent" because it is composed by several specific extractors for each elements of the image without any dependence between them. Extra processing such as balloon type classification and text recognition are also covered. The third approach introduces a knowledge-driven and scalable system of comics image understanding. This system called "expert system" is composed by an inference engine and two models, one for comics domain and another one for image processing, stored in an ontology. This expert system combines the benefits of the two first approaches and enables high level semantic description such as the reading order of panels and text, the relations between the speech balloons and their speakers and the comic character identification.

Keywords: image processing, graphics recognition, document analysis, comics understanding.

Laboratoire L3i - Informatique, Image, Interaction

Pôle Sciences et Technologies, Université de La Rochelle, avenue Michel Crépeau

17042 La Rochelle - Cedex 01 - France