International Journal on Document Analysis and Recognition manuscript No.
(will be inserted by the editor)

Knowledge-driven understanding of images in comic books

Christophe Rigaud[1,2] · **Clément Guérin**[1] · **Dimosthenis Karatzas**[2] · **Jean-Christophe Burie**[1] · **Jean-Marc Ogier**[1]

Received: date / Accepted: date

Abstract Document 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 identify the key elements of a comic book story and arrange them according to a predefined domain knowledge. In this study we propose a knowledge-driven system that can interact with bottom-up and top-down information to progressively understand the content of a document. We model the comic book's and the image processing domains knowledge for information consistency analysis. In addition, different image processing methods are improved or developed to extract panels, balloons, tails, texts, comic characters and their semantic relations in an unsupervised way.

Keywords document understanding · comics analysis · expert system

1 Introduction

Comics or "bandes dessinées" represent an important part of the cultural heritage of many countries, especially in the US [52,24], Western Europe (particularly France and Belgium) [41], and Japan [3]. Unfortunately, they have not yet received the same level of attention as music, cinema or literature in terms of their adaptation to the digital format. Using information technology with classic comics would facilitate the exploration of digital libraries [4], assist translators [6], provide a tool for augmented reading [49,26], speech playback for the visually impaired [7,39], 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. It is defined as juxtaposed sequences of image by McCloud [34] and Thomas [59]. The frame of the page, its size and its organization matter to the author who can use them to tell his story the way he wants to. As opposed to books or movies, changing the medium of a comic books might change its artistic dimension, resulting in a potential betrayal of the author's original intention.

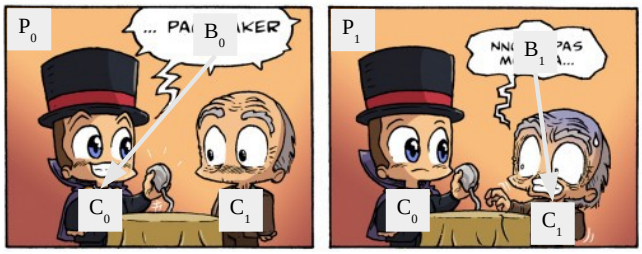


Fig. 1: The left panel P_0 represents a comics character labelled as C_0 saying the content of balloon B_0 to another character labelled as C_1. In the right panel P_1, the character C_1 is saying B_1 to C_0.

Ideally, based on an understanding of the process used by the authors to draw the paper version, the comics would automatically be changed into a form adapted to the medium in which the work is to be read (e.g. smartphone, web page, 3D book). This would begin with an analysis of the digitized paper page to extract the different content elements. Then, the story would be reconstructed by placing the extracted elements in the narrative order, through an understanding of the role of each panel in the story, and the association of

[1] Laboratoire L3i
Université de La Rochelle
17042 La Rochelle CEDEX 1, France
E-mail: {christophe.rigaud, clement.guerin, jean-christophe.burie, jean-marc.ogier}@univ-lr.fr
.
[2] Computer Vision Center
Universitat Autònoma de Barcelona
E-08193 Bellaterra (Barcelona), Spain
E-mail: dimos@cvc.uab.es

speech balloons and the comic characters (protagonist)[1] to whom they are addressed, etc. (Figure 1).

In this paper we address the problem of understanding comics content by automatically improving the consistency of the extracted elements according to the domain knowledge and by inferring new information and launching further processing iteratively. Low level and high level information is used in a complementary and unsupervised way. We progressively identify the different elements of a digitized comic book image and retrieve their relations. To do so, we conceptualized and formalized the underlying semantics related to their positions in the image. Another novelty lies in our proposal for improving the state of the art methods for panel, balloon, tail, text and comic character detection and description to take advantage of this framework.

The proposed framework is based on interactions between low and high level processing in order to reduce the challenging semantic gap. In this paper, we call each image processing algorithm an "extractor" of panels, balloons (or bubbles), tails and comic characters according to the type of region they are designed for. Extractors provide a set of regions that feeds a knowledge base. The knowledge base is associated with rules that form an ontology. 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).

Fig. 2: Examples of comic panels that reflect the diversity of comics (Golden age American comics, old European comics, recent European webcomics and recent manga).

[1] In the rest of the paper "character" is used in the sense of actor or protagonist of a comic's story, not as a piece of text.

What we called our "expert system" uses the ontology to assert the relations between regions, according to the formalized knowledge. In addition, an inference engine is used to deduce new pieces of information in order to perform further image processing (e.g. hypotheses of the regions of interest for the localisation of comic characters or speech balloons). Please note that the term "expert system" is used here to name our global reasoning software, not as a reference to the rules-based deduction systems developed a few decades ago in the literature [17].

The rest of the paper is organized as follows. The related work for holistic document understanding and comic book analysis techniques is summarized Section 2. The proposed framework is described in Section 3. Subsequently, the models, the low level processing and the interactions between them are detailed in Section 4, 5 and 6 respectively. The datasets used to evaluate our method, the evaluation protocols and the final results are shown in Section 7. Finally, results are discussed and conclusions drawn in Section 8.

2 Related work

In this section we provide an overview of work on the holistic understanding of documents and then provide the state-of-the-art studies related to an analysis of the literature on comic book images.

2.1 Holistic document understanding

One of the original goals of image document analysis was to fully understand the content of any images [29]. The meaning of the term "understand", when it comes to machines, can be taken as their ability to recognize graphical elements and to interpret the semantic of their features. This requires solving several sub-tasks simultaneously, for instance region detection, labelling of meaningful regions and semantic interpretation using layout analysis. In the past, researchers have often developed classifiers for tackling each of these sub-tasks independently [33]. However, these sub-tasks can help each other. For instance, in a comic book page, if we know the panel positions, then we can make a better guess about the location of the comic characters (they are usually in the panels). It is not easy to combine different related sub-tasks. Previous work concerned real scene image analysis [5], retrieval [8] and understanding [30,12], medical image annotation using description logic and inference engine [22], object-based image retrieval [36,48] (between keyword-based and query-by-example) and image interpretation [23,37]. Most researchers agree on the importance of using the underlying semantic of topological information, spatial distances, directional relative position and complex relations such as "between", "surround" and "among", to

obtain an exhaustive description list of the contents for a given context.

Recently, images from comic books have been also considered. In [18], a semantic annotation tool made use of previous knowledge and consistency information to suggest new knowledge to the user in an interactive way. Spatial inferences have been used to infer the reading order of comic books, at the level of panels in each page and balloon in each panel [13]. The benefit of using contextual information of a simple object to build more complex ones has been highlighted [8]. To our knowledge, there is no framework for comic book images understanding in the literature that infers new knowledge iteratively and without user interaction (unsupervised).

2.2 Analysis of comics

Images in comics are mixed content documents that are processed differently depending on the type of content that we are interested in. The techniques involved can vary a lot depending on whether we focus on panels, balloons, text or processing comic characters. We will review each of these in the next four paragraphs.

Panels Panel extraction and ordering has mainly been studied for panel to panel reading. This kind of requirement has been continuously increasing in parallel with the evolution of mobile devices. Readers want to have their favourite comics or mangas on the go, while carrying minimum weight. Printed comics need to be manually scanned and split into screen size parts small enough to avoid zooming and scrolling, which is tedious.

Several techniques have been developed to automatically extract panels [25], assuming that panels are small enough elements to be comfortably read on mobile devices. Most of them are based on white line cutting with Hough transform [10,31], recursive X-Y cut [16] or density gradient [58]. These methods do not consider empty area [25] and border free panel. These issues have been corrected by connected component approaches but these approaches are sensitive to regions that sometimes connect several panels, which potentially increases the detection error rate [1,46]. Another approach based on morphological analysis and region growing can remove such connecting elements but at the same time creates new holes in the panel border [19]. After region segmentation, heuristic filtering is often applied to classify panel region according to their size ratio with the page size [2,19]. More recently, new methods have shown interesting results for manga and European comics with different background colours. They are based on watershed [40], line segmentation using Canny operator and polygon detection [31], and region of interest detection [53] such as corners and line segments.

Balloons Balloons or bubbles are key elements in comics, they link graphic and textual elements and are part of the style of comics. They can have various shapes (e.g. oval, rectangular) and contours (e.g. smooth, wavy, spiky, absent). Closed speech balloon have been studied in the main, based on region detection and filtering rules [2,20]. Our group developed a method to extract open (non closed) balloons based on active contours initialized around text areas [44] and an initial approach to classify balloons according to their contour styles [42].

A balloon is a spatial container of information that is related to a protagonist using a specific element: the tail. The tail is often represented by a discontinuity on the contour of the balloon towards the concerned protagonist. To our knowledge, there is no work on tail position and direction detection in the literature but we have stressed its interest for improving balloon classification [42].

Text Text extraction and recognition have attracted a lot of attention in complex analysis domains such as real scenes, cheques, maps, floor plans and engineering drawings. However, few contributions concern comics. Bottom-up approaches use connected components which often rely on a segmentation step [40]. Su used Sliding Concentric Windows for text/graphic separation and then used mathematical morphology and an SVM classifier to classify text from non-text components [54].

We have previously developed an adaptive binarisation process based on Minimum Connected Component Thresholding followed by a text/graphic separation based on contrast ratio and text line grouping [45]. Li proposed an unsupervised speech text localization for comics based on training a Bayesian classifier on aligned connected component and then detecting the rest of the text using the classifier for text/non text separation [32].

Top-down approaches starting from speech balloon (white blob) detection followed by mathematical morphology operations have been proposed by Arai [2] and layout analysis by Yamada [61].

Comic characters Human body and face detection in real scene images has progressed a lot during the last decade. Nevertheless, the few studies that are related to comics showed that human detection techniques are not appropriate for comic book characters. Those studies concerned exact and partial copies of mangas [55], cartoon classification [27] and the detection of the main characters using redundancy information [56,21]. We have recently developed a query by example approach based on a colour descriptor [43].

3 Framework for understanding comic books

In this section, we introduce an unsupervised framework for understanding comic book images, based on our expert system combining a knowledge base and reasoning capabilities. We also detail its application in a concrete situation.

3.1 Expert system

The purpose of our expert system is to interact with the low level (image processing) iteratively to progressively understand the content of an image, moving from simple to more complex elements. This approach is similar to [8] except that in our case the definition of the complex object is not a composition of simple objects but context-driven.

In our system, the expert system includes two models, one formalizing the raw data from algorithms (*Image model*) and the other modelling the domain knowledge of the comic books (*Comics model*). These two models are ontologies that work together to express the relations between the primary elements of a comic book image that can be considered as being stable through all instances of the studied domain (Figure 3). Thus, the expression of the constraints applied both to the elements and their relations have to be specific enough because these constraints will be considered as the reference knowledge for the detection of potential errors of the low-level extraction algorithm output. The main advantage of separating image knowledge from the application domain knowledge is that it makes this framework more easily adaptable to other applications by replacing the second ontology only.

The low level algorithms are 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.

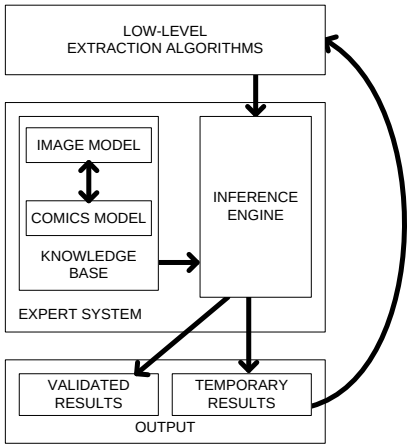


Fig. 3: Generic representation of the expert system and the relationship between knowledge base, the inference engine and the low-level algorithms.

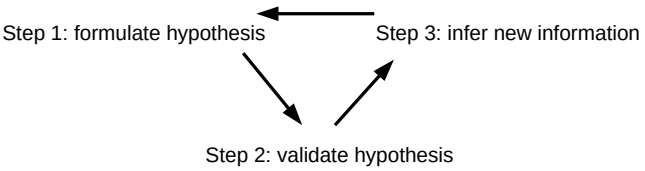


Fig. 4: Process loop of the framework.

In Figure 4, the starting point is step 1 (formulate hypothesis) where low level processing give basic information to the expert system. Then, step 2 (validate hypothesis) assesses the valid elements and removes obvious mistakes, and in step 3 we infer new information based on the previous information and the knowledge base. In the next iteration of the process loop, the newly validated information can be used by low level processing (e.g. new parameters, region of interest) to extract more complex elements from the image and so on. This loop can potentially be run as many times as new information is discovered, depending on the application and the amount of information to be retrieved. In our case, the process is made of two iterations as described in Section 3.3.

3.2 Knowledge representation

The domain knowledge is conceptualized though the categorization of each element composing a page, combined with a set of topological relation 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 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, American). We conceptualized the general properties of a comic book image layout 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 same character can appear only once in a panel

– A text line T is related to one and only one balloon B

Despite the fact that authors are entirely free in their layout choices, some researchers insist a few conventions, widely adopted by comic book' authors, be respected to avoid the reader being confused [28, 9]. The depicted elements and their place in the layout must be clearly identifiable at first sight, meaning, for instance, that balloons and characters should be included inside panels. Whereas one can find some instances of balloons breaking out of their frame, 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). 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 [14]. Figure 5 shows the percentage of panels, balloons, text lines and characters that fit the enumerated constraints, as a function of their covered area.

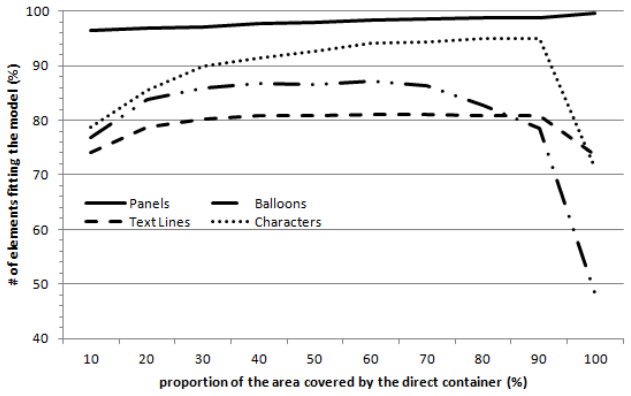


Fig. 5: Percentage of panels, balloons, text lines and characters from the eBDtheque dataset [14] that fit the definition given in 3.2, depending on the area that need to be covered.

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. The lowest score was obtained for the text lines and was predictable as our corpus integrates ten pages with old-fashioned text captions, the concept of speech-balloon being not invented yet at the time of their publication.

Then, we introduced some refinements for balloons, text lines and characters, respectively named speech balloon SB, speech text ST and speaking characters SC. Their definition is based upon the formalization of the semantic carried by the following spatial configurations:

– A SB is a balloon B that has a tail and contains text
– A SC is a character C pointed by a tail
– A ST is a text line which is included in one speech balloon

The term "pointed" refers to the fact that the character is included in the part of the panel indicated by the tail. To have a better understanding of how a panel is divided according to tail direction, please refer to Section 6.2.

These figures represent the theoretical limit that can be reached with the model in terms of component extraction performance.

3.3 Processing sequence

The expert system asserts the extraction of simple elements such as panels, texts, balloons and tails in order to infer speech balloons before searching for comic book characters (which are much more complex), based on the context defined by the simple elements and their relations. This can be achieved through the two iterations of the process loop as shown in Figure 4. These sections are detailed below.

Iteration 1 - step 1 (hypothesis) The initial extraction of panels, text and balloons feeds the knowledge base. In Figure 6, dashed elements represent the initial hypotheses. Note that extraction errors can take place at this stage which the system can recover from at a later stage.

Iteration 1 - step 2 (validation) Subsequently, the expert system checks if their spatial relations (context) match the topological properties of the knowledge base defined in section 3.2, otherwise it solves them using the domain knowledge described in Section 6.1. The result is illustrated in Figure 7.

Iteration 1 - step 3 (inference) From the validated information, the expert system infers the specification of some balloons and text lines into speech balloons and speech text lines, with respect to the formalized semantic in the knowledge base (Figure 8).

Iteration 2 - step 1 (hypothesis) This step is the beginning of the second iteration of the process illustrated Figure 4. At this point the expert system already has some information about the content of the image from which a further hypothesis can be made concerning complex elements such as the region of interest (ROI) of the characters from the speech

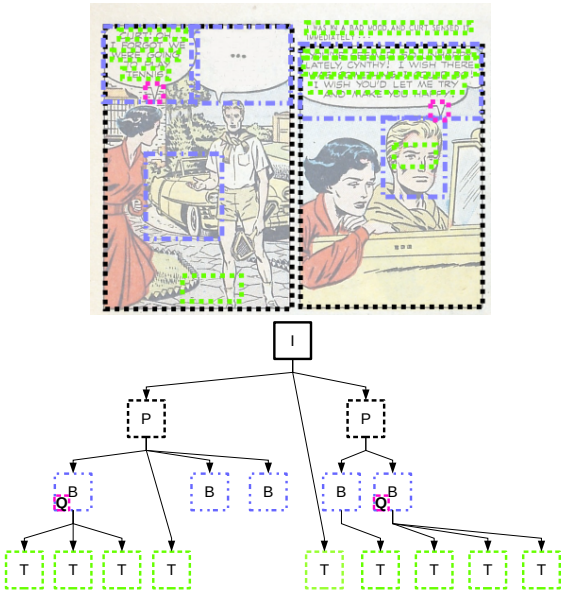


Fig. 6: Initial hypothesis (dashed elements) about the content of a given image I after the initial extractions of panels P, text T and balloons B with tails Q.

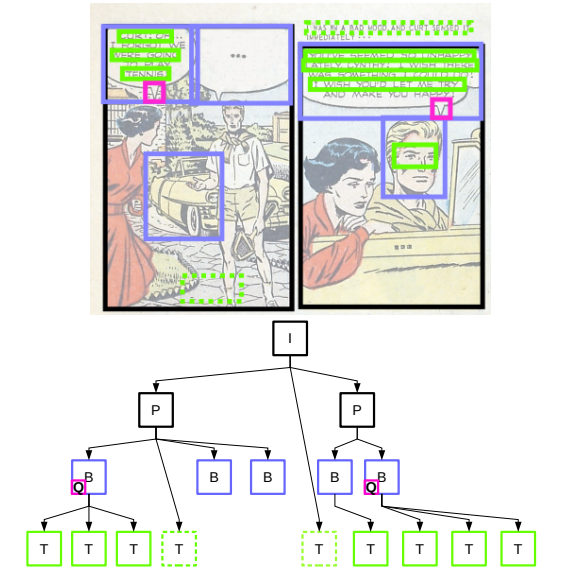


Fig. 7: Validation of the hypothesis using the properties of the knowledge base. Valid elements have a solid border.

balloon positions (Figure 9). The ROIs defined by the expert system are given as seeds to the image processing algorithm (extractor of characters) which in turn feeds the expert system with more precise character locations (Figure 10).

Iteration 2 - step 2 (validation) The expert system checks if the spatial relations of the characters C match the properties of the knowledge base defined in Section 3.2 (Figure 11).

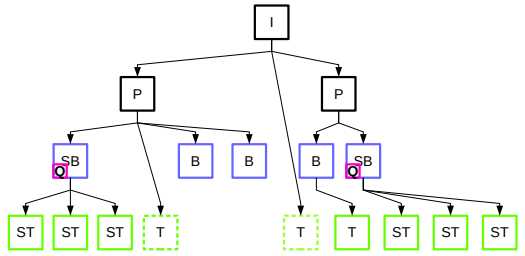


Fig. 8: Inference of the speech balloon SB and speech text ST regions using the semantic properties of the knowledge base.

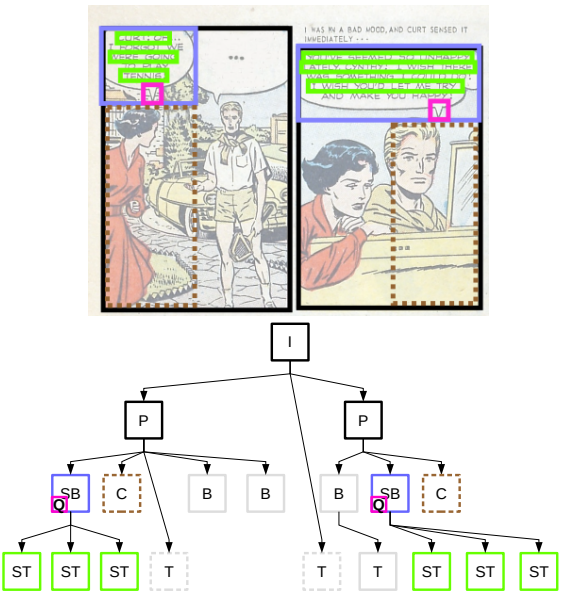


Fig. 9: Hypothesis of ROIs of characters C from the speech balloon SB regions and the corresponding image I. The regions that are not related to ST have been shaded in the graph and removed from the image to make it more comprehensible.

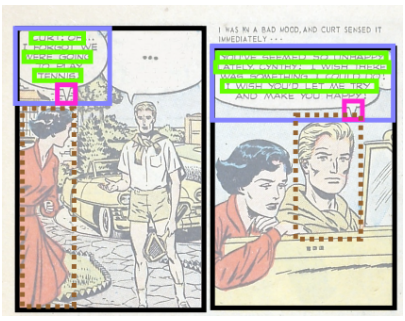


Fig. 10: Character locations (C) returned by the low level processing from the ROIs defined in Figure 9.

Iteration 2 - step 3 (inference) The expert system infers which characters are speaking SC (Section 6.2) and link them to the corresponding speech balloons which have already been

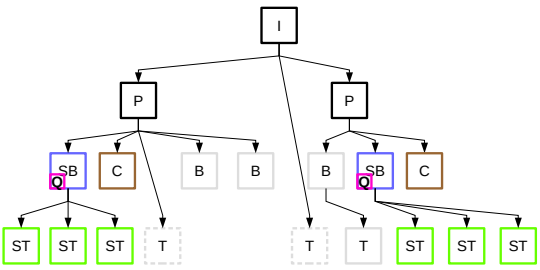


Fig. 11: Validation of the character regions (C) by the expert system and the corresponding image I.

linked to speech text regions in *Iteration 1 - inference* step. (Figure 12).

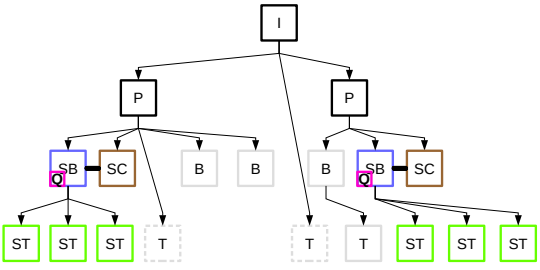


Fig. 12: Inference of the two speaking characters (SC) and the corresponding relations between speaking characters SC and speech balloons SB regions. Both links are represented by a grey stroke in the image over the regions concerned and a non oriented horizontal edge in the graph.

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

4 Knowledge formalization

Our knowledge base is composed of two ontologies, designed using OWL's W3C recommendation [35], and interacting with each other (Figure 3). The first one was used to model the raw data provided by image analysis algorithms (called *image model* hereafter), while the second formalize the conceptualization of the comic book domain knowledge (called *comics model*), described in section 3.2. These two models are bounded by bridges that are used to perform reasoning over both models, using their own properties.

The comics model formalizes the introduced elements (panels, balloons, text lines and characters) into OWL concepts. The relations (affiliation and association) between them are represented by object properties. Their respective *range* and *domain* are set accordingly to the constraints introduced in the layout conceptualization to detect potential extraction errors. Additionally, the concepts of "Speech Balloon", "Speech Text" and "Speaking Character" are derived from their respective super-concepts via an OWL translation of the rules we defined earlier.

The image model is for the main part composed of the concepts of *Image*, *Region Of Interest (ROI)* and *Extractor*. A region of interest being a part of image identified by an extractor as a relevant piece of content for some application. In the present case, an image is identified as a comic book's page and a region of interest as one piece of content (e.g. a panel). This equivalence is translated through the OWL axiom "equivalent class". This provides a way to handle comics elements as regions in the image and allows reasoning processes on their shape and position.

5 Low level processing

Low level processing is used to extract information from the image without taking into account the context unless it is given from high level processing. Panel, balloon, text and comic character extractions we used are methods from the literature that we improved or adapted for this work. As far as we know, tail detection have not been studied before. Here we propose an initial method based on balloon contour analysis.

5.1 Panel extraction

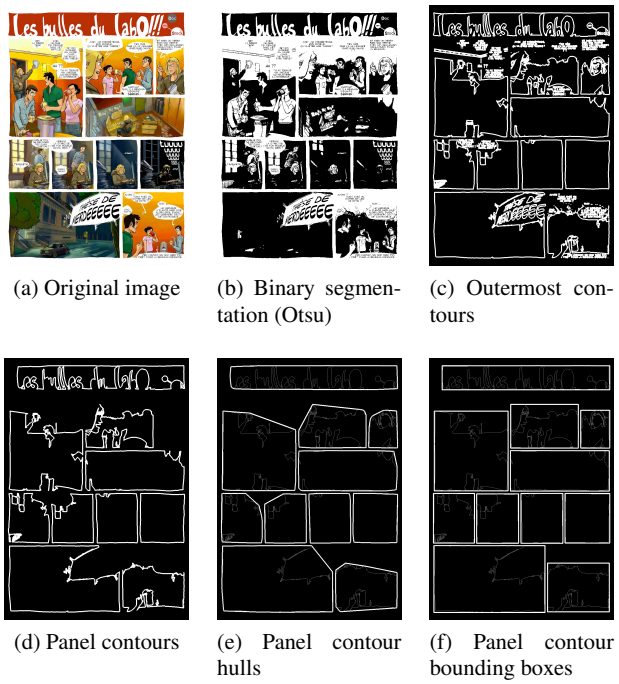
We propose to combine the advantages of the already published panel extraction methods [2,46] and partially solve their weaknesses using the expert system model. They are connected component based methods that are reliable for comics with disconnected panels (separated by gutters). Line-based decomposition methods are the most efficient methods for line separated comics [32] (no gutter). A generic method that works best with both styles is not easy to implement. In our system, we combine [2,46] to introduce a new method that is especially suitable for general comics using gutter (white space) between panels.

In [2], the authors considered panels as being connected components (closed frame). They propose to extract black connected components from a binary image and use heuristic filters to prune non-panel candidates. In their approach, the binary image is produced from a fixed thresholding of the grey version of the image. In [46], we have showed the benefits of using a global and adaptive thresholding method (like Otsu's approach) to extract connected components corresponding to panels. In this approach, we clustered the components into panel (big), text (medium) and noise (small) but the clusters are inter-dependents.

So we reused global thresholding [46] to extract connected component similarly to Arai [2] and propose a novel topology-based filtering process to retrieve connected components corresponding panels. The global thresholding method separates page background, expected to be clear, from its content considered as foreground here (dark elements such as black strokes), see Figure 13b.

Note that local thresholding approach is not used here because it has higher chances to split the panel border and over-segment its content. Then, the connected components from the binary image are extracted and their topological relations are analysed to filter out non-panel components [57]. One of the particularities of the panels is that they are usually not included in other elements. Therefore, we focus only on the external (outermost) connected components (Figure 13c). Note that we filter out small components in order to avoid considering isolated elements (e.g text, logo, page number) as panel, (Figure 13d). Finally, the convex hull or the bounding box of the panel contours can be computed in order to recover from discontinuous contour detection (Figure 13e and 13f).

The evaluation of the proposed method is shown in Section 7.1.



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

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

Fig. 13: Panel extraction process.

5.2 Balloon extraction

Of the three methods in the literature, two of them are based on text position which relies on the text extractor performance [20,44]. We didn't use these methods because we preferred to process low level information independently in order to avoid error propagation. Arai's method does not use text position but connected component filtering based on heuristics (fixed thresholds for binarisation, blob size, number of white pixels, number of contained straight lines and width to length ratio) [2]. We developed a more adaptive method that can handle various styles of comics (e.g. colour, B&W), image formats (e.g. A4, A5, portrait, landscape) and digitization variations (e.g. scanning device, camera capture). Fixed thresholding such as those used in [2] are a strong assumption and only work for comics with the same style and digitized under the same light condition as already mentioned in Section 5.1. We relaxed this constraint by using the Otsu threshold selection method to make a binary segmentation of the greyscale image [38]. Binary segmentation creates a lot of small connected components on textured images. This was pruned using [46] (minimum average height class). We then analysed their parent-child relationship to make a first selection of candidate components before analysing the spatial organisation of the set of children. We defined the set of children $CH = \{ch_1, ch_2, ..., ch_n\}$ and compute a confidence value for each parent from the set $P = \{p_1, p_2, ..., p_m\}$ that serves as final filtering criterion. The process is illustrated in Figure 14.

We assume that a balloon has a minimum of children $minNbChildren$ which are horizontally or vertically aligned and centred inside the balloon region (property of text in comics). The percentage of alignment $align$ is computed for

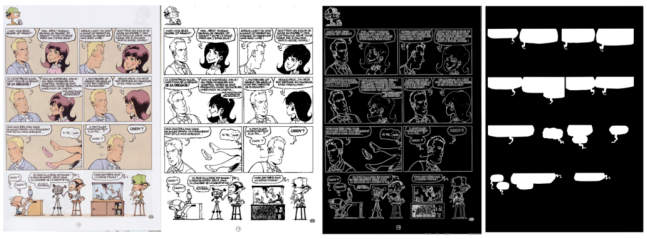


Fig. 14: Balloon extraction process. Original image, binarisation (from greyscale image), contour detection (connected components) and result mask from left to right.

each child ch_i considering $CHA(ch_i)$ the subset of children that are aligned to ch_i, Equation 1.

$$align(ch_i) = \frac{|CHA(ch_i)|}{n} \quad (1)$$

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.

The difference of coaxiality is computed for the horizontal (dx) and the vertical (dy) axis from the Euclidean distance between the parent and children hull centroids (Figure 15). Both differences dx and dy are normalized between

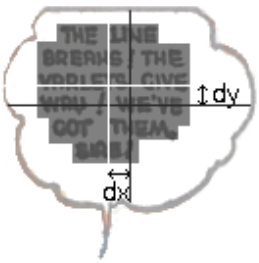


Fig. 15: Horizontal and vertical coaxiality measurement between the centroids of the balloon (dark cross) and the children hull (clear cross). The children hull is the grey region included in the balloon.

zero and one as a percentage of the balloon (parent) width and height respectively. The average of the alignments, $\overline{align}$, dx and dy gives a confidence value for each candidate of P (Formula 2).

$$C_{balloon} = \frac{1}{3} * \left(\overline{align} + dx + dy\right) \quad (2)$$

The confidence value was used for balloon and non balloon separation in the experiment Section 7.3.

5.3 Tail detection

Speech balloons indicate dialogue, sometimes using a tail pointing at their respective speakers [47].The tail is usually represented by a discontinuity on the balloon contour (Figure 16). We proposed an initial tail extractor of tail tip (extremity) and tail direction detection based on balloon contour analysis. We limited this study to the tail types that can be considered as an extension of the inside region (background) of the speech balloon namely “comma”, “zigzag” and “absent” (no tail) because they can all be extracted from the segmentation of the balloon background (Figure 16). Other types require a specific study based on the speech balloon extraction and considering contour strokes and surrounding elements. Note that tail direction was named after the eight cardinal directions (North, North Est, Est, South Est, South, South West, West and North West). This facilitated further processing and evaluation.

A tail can be decomposed into four elements: origin, path, tip and pointed direction. The contour of a speech balloon is mainly convex, except in the region where the tail is connected to the balloon (origin) which produces the highest convexity defects. If the speech balloon has no tail, there is no particular convexity defect. We based our approach on the analysis of this specific region (where the highest convexity defects are). 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 17). 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 defined the top two farthest points f_a and f_b corresponding to the convex hull segments s_a and s_b, as the coordinates of the tail origin (Figure 17).

Most of the time, the tail tip corresponds to one of the vertices of the convex hull. We defined the set of vertices $V = \{v_1, v_2, ..., v_n\}$. The optimal vertex was computed by comparing five features that correspond to the Euclidean distances:

- ds_a the distance to the segment s_a
- ds_b the distance to the segment s_b

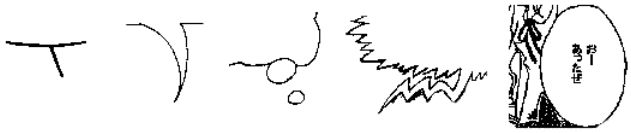


Fig. 16: Examples of type of speech balloon tails. From left to right: stroke, comma, circle, zigzag, absent.

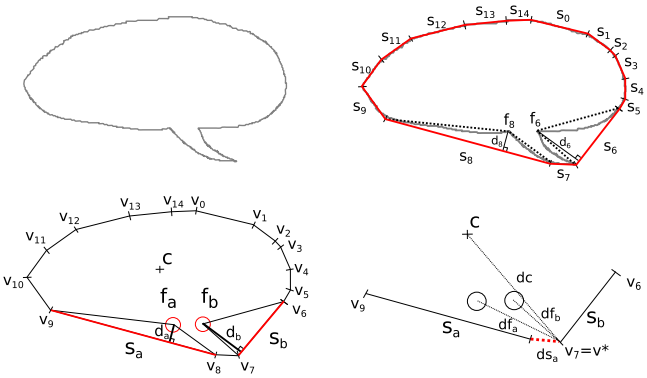


Fig. 17: Convex hull and convexity defects of a speech balloon (top part). Bottom-left figure represents the two biggest convexity defects of depth d_a and d_b which define the tail origin between the two corresponding points on the balloon contour f_a and f_b. Bottom-right figure represents the distances related to the vertex v_7 to illustrate the variables of Equation 3. In this case $ds_b = 0$ because v_7 coincides with an end of segment s_b (distance equal to zero).

- 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:

$$v* = argmax\big(max(dc+df_a+df_b)+min(ds_a+ds_b)\big) \quad (3)$$

where $v*$ is the optimal vertex from the set of vertices V.

Together with the position, we compute a confidence value C_{tail} related to the mean depth of f_a and f_b over the mean balloon size $meanBalloonSize$ (Formula 4).

$$C_{tail} = \frac{(depth(f_a) + depth(f_b))/2}{meanBalloonSize} \quad (4)$$

where $depth(f_x)$ is the depth of the corresponding defects (e.g. d_a and d_b in Figure 17) and $meanBalloonSize$ is defined in Equation 5.

$$meanBalloonSize = \frac{\sum_{i=0}^{n} Wb_i + \sum_{i=0}^{n} Hb_i}{n * 2} \quad (5)$$

where Wb_i and Hb_i correspond to the width and height of a speech balloon i and n the number of speech balloons in the image.

Once we found the tail tip, we analysed its neighbourhood to find the orientation and direction of the last part of the tail which is directed towards the speaking character. We computed a MxM square region centred on the tail tip position in the balloon mask and used a line fitting algorithm (weighted least-squares) that returns the representative line

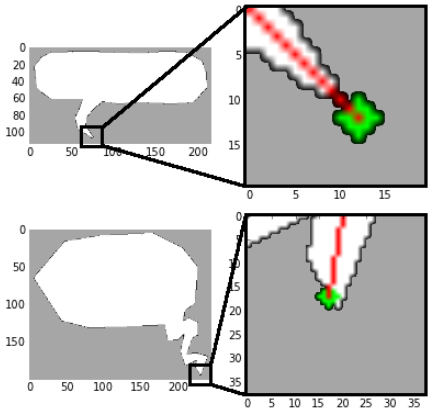


Fig. 18: Tail orientation given by the line fitting algorithm in the neighbourhood of the tail tip. Tail tip is represented as a green square and the result of the fitting algorithm by a red straight line.

that minimizes the distance between all the points of the local region (Figure 18).

The tail direction was computed from the tail orientation that has initially two possible directions. We defined the tail direction from the closest point of the representative line to the tail origin called $fitPt$, to the tail tip ($\overrightarrow{fitPt, tailTip}$).

5.4 Text extraction

In our method, the expert system uses the text information to deduce which balloon is actually a speech balloon (Section 3.2). Several text extractors have been reported in the literature, both for Latin and syllabic scripts. We used our own method because its ability to handle a large variety of comics has already been demonstrated [45]. In this work, although our aim was not to recognize text, we nevertheless used an Optical Character Recognition (OCR) system to improve the quality of text localization by filtering out regions where no alphanumeric symbols were recognised. The transcription obtained was stored in the ontology for future processing. Instead of training an OCR for a specific language and font like [40], we preferred to use Tesseract OCR which has already been trained with several fonts and languages [51].

5.5 Character extraction

Unsupervised comic book character extraction becomes a very difficult task when processing heterogeneous comic styles such those in the eBDtheque [14] dataset (Figure 2). In this context, learning-based approaches [27,55] can not handle such dimensionality induced by the difference of styles. We

(a) Inferred ROI (b) User defined ROIs [43]

Fig. 19: Comparison of comic character region computation, automatically inferred (left) and user defined (right).

previously developed a method that learns a specific character based on a single user defined example [43]. In this study we decided to go one step further by detecting comic character positions without requiring user interaction (unsupervised). We propose to ask the expert system to "guess" a ROI where comic characters might be according to the other elements in the panel (ignoring image features) and then apply low level processing to refine the ROI using image features (Figure 19).

As introduced in Section 6.3, the hypotheses of the ROI of comic characters were given by the expert system using a combination of low and high level information. The role of the comic character extractor here is to refine the ROI proposed by the expert system in order to better fit the position of the character in the image (using image features). This is an optimisation problem where we want to maximise the ratio of pixels that belong to the comic character over those that belong to the background in the guessed ROI (from expert system). At this point, deciding whether or not a pixel is part of a character is difficult. Instead, we preferred to work at region level using closed regions (regions fully surrounded by a black stoke) and deciding whether or not a region belongs to the character. We considered a region to be part of the character if most of its pixels were contained by the predicted ROI. This method excludes wide region pixels where most of the pixels are out of the ROI such as background regions and completely includes small regions that were partially included. It can be applied to most of the comic book styles and works best when the characters are entirely overlapped by the ROI hypothesis with a uniform background (Figure 20). Note that speech balloon and text regions were ignored because we did not expect a character to intersect with those regions as discussed in 3.2.

6 Interactions between high and low level processing

This section presents the low and high level processing interaction to validate, infer information and generate hypotheses about image content.

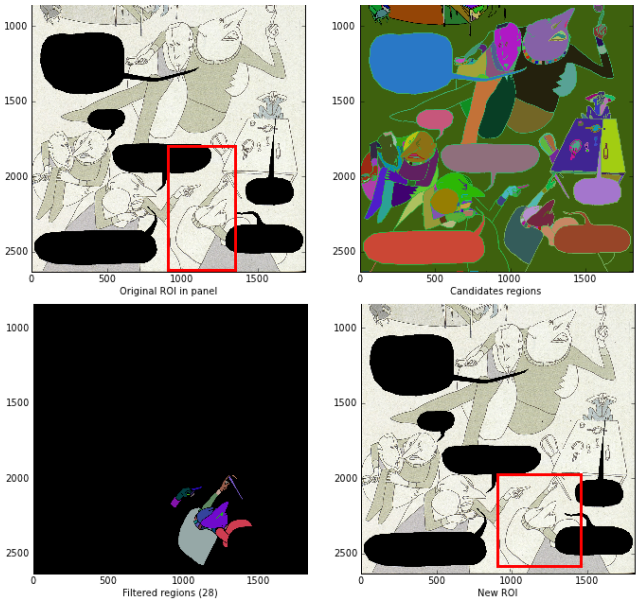


Fig. 20: Hypothetical ROI (red rectangle), region segmentation, region selection and character extraction bounding box (red rectangle) from top-left to bottom-right.

6.1 Validation of the extractions

In order to validate the extraction of the comic page's components, we made sure that the extracted panels, balloons and text lines were in accordance with the knowledge conceptualized in Section 3.2.

Firstly, each item was loaded into the model as it was labelled (page, panel, balloon, text line or comics character) by the low level processing step. Then, each element was linked to its smallest container (as defined in Section 3.2) in a half-blind way. That is to say that the type of contained element was known, while the type of potential container was not. Not knowing the type of container might have led to incorrect assertions that would have produced inconsistencies in the model. Those inconsistencies are the result of possible mistakes made during the extraction process that were filtered out in order to improve the overall detection precision. Consistency checking was performed over the model and inconsistencies were handled one after the other. We observed that the misclassification of elements (e.g. a balloon labelled as a panel) was not the greatest weakness of the current algorithms. Therefore, we chose to focus on increasing the extraction precision by filtering out the elements that did not fit the constrained model. This caused an inevitable drop of recall that will be addressed in the upcoming works. For the time being, our system can handled, without being limited to, the following inconsistencies:

– **A page (p) contains a balloon (b), a text line (t) or a character (c)**: b, t or c 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 balloon (b) contains a character (c)**: c is deleted.
- **A text line (t) contains a panel (p), a balloon (b) or a character (c)**: if p, b or c does not contain any legitimate elements then p, b or c 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.
- **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.

6.2 Inferences from the low level information

The expert system is able to infer more specific information than that given by the extractors. For instance, it can deduce which text is spoken or not and which speaking character pronounces the content of which speech balloon.

Speech balloon and speech text The classification of balloons and text respectively into speech balloons and spoken text can be done by running an inference engine on the model (e.g. Racer [15]] or Pellet [50]). To allow reasoning, the model has to be consistent, which is the case after validation step 6.1 when all inconsistencies have been resolved. In Section 3.2, we defined a speech balloon as a balloon that has a tail. In others words, the concept of a *speech balloon* can be seen as a specialization of the *balloon* concept that extends all its properties plus adding a few new ones. This is expressed in the expert system by defining a property *hasTail* with the constraint that must have a *speech balloon* instance as a source. Assessing this piece of knowledge, the reasoner will automatically deduce that each instance of *balloon* extended with a *hasTail* property can be specialized into a *speech balloon*.

In a similar way, the concept of *text line* subsumes the concept of *spoken text line*. It is rendered equivalent to the set of individuals from the *text line* concept that is bound with the property *isLineOf*, which is the inverse property of *hasLine*, to a *speech balloon*. In other words, the text lines marked as being part of a speech balloon are automatically classified as spoken text lines.

Speaking characters Among the validated characters (Section 6.1), we considered as being potential speakers those who intersect with the hypothetical region computed in Section 6.3. These regions were computed from each speech balloon, i.e. the balloons that had an identified tail. An abstract straight line was drawn from that tail tip in the direction indicated by the tail. The first region of a potential speaker that it touched was considered to be the source of the speech balloon. This relation was asserted into the ontology with the property *isSaidBy*, between the selected character and the corresponding balloon. Since the range of this property was set to the concept of *speaking character*, it automatically classified the character instance involved into this class.

6.3 Hypothesis of ROI of characters

To detect comic characters, we started by narrowing down the area where we were going to look for them. This could be done using the structured information built up on the image during the previous steps. We focused on the characters emitting a speech balloon (speaking characters) and tried to localize them from the speech balloon informations. That is to say that speech balloons are associated once or several times to each of the characters in different panels.

As we knew which balloons were speech balloons and what panels they appeared in, we could estimate, from the position and direction of the tail (Section 5.3), which part of the panel contained the speaking character. Tail orientation was quantized in eight cones of $\pi/4$ radius so we can consider that, in a rectangular-shaped panel (or its bounding box), the tail was either pointing towards a corner or towards a panel border. We defined the ROI for the character as a squared region beside the speech balloon. We defined the maximum width w_{max} and height h_{max} of the ROI equal to the mean widths and heights of all the balloons in the image (Equation 5).

The position of the ROI around the speech balloon was defined by two opposite points of a square $A_{x,y}$ and $B_{x,y}$ according to Formula 6 which is sometimes constrained by a panel border which is why the term $min()$ appears. The different positions of the ROI are illustrated on Figure 21.

$$\begin{array}{ll} A_x = & v_x^* - min(Pi_x - v_x^*, w_{max}) \quad *O_x \\ A_y = & v_y^* - min(Pi_y - v_y^*, h_{max}) \quad *O_y \\ B_x = & [A_x + min(Pi_x - A_x, w_{max})] *O_x \\ B_y = & [A_y + min(Pi_y - A_y, h_{max})] *O_y \end{array} \tag{6}$$

where v^* is the coordinates of the tail tip (Section 5.3), Pi the coordinates of one of the four corners of the panel bounding box $P = \{P0, P1, P2, P3\}$ and O the offset for the direction of the tail. The offset O was quantized in heigh values according to the tail direction and the panel corner P_i was chosen accordingly (Table 1).

The ROI were then passed to the comic character extractor (Section 5.5) which will locate more precisely the characters based on image features.

Table 1: Values of the offset O on the horizontal and vertical axes and panel' 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

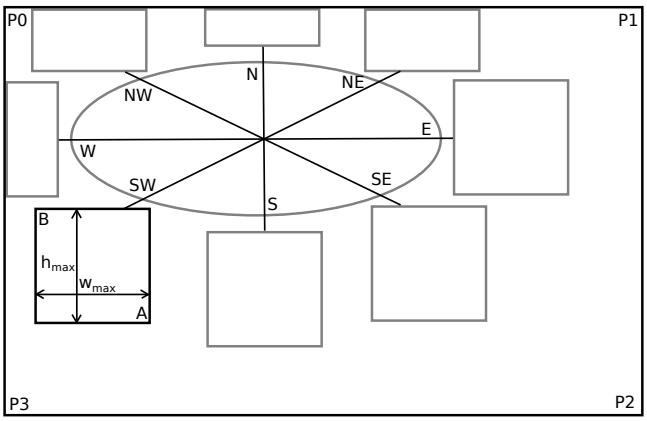


Fig. 21: Illustration of a panel with four corners $P0, P1, P2, P3$ and the different ROI positions and sizes that can be defined for each of the eight directions of the tail.

7 Evaluations

In this section, we describe the dataset, the ground truth and the metrics we used to evaluate our work. At first we focus on each low level image analysis contribution independently in order to avoid error propagation effects. Secondly, we evaluated the general framework combining the different contributions of this paper (low and high level processing). Results and complementary materials are available online for each image of the dataset[2].

Dataset and ground truth We evaluated our different contributions using the public dataset eBDtheque [14] of one hundred comic book images with the second version of the ground truth "version 2014" that contains 850 panels, 1550 comics characters, 1092 balloons and 4691 text lines. This dataset contained 46% of images scanned from French comic books at 300 DPI (dots per inch), 37% of images from French webcomics with various formats and definitions, 11% of public domain American comics[3] in A4 format and 300 DPI and 6% of unpublished artwork of Japanese manga in A5 format at 72 DPI. Five percent of the images contains a double page. In addition to the diversity of styles, format and definition, there are also differences in design and printing techniques since 29% of the images were published before 1953 and 71% after 2000.

In the second version of the ground truth, the main novelty compared to the first version concerns the balloons and the comic book characters. The balloons were segmented at pixel level on the external edge of their contour (previously at bounding box level) in order to evaluate their extraction and shape more precisely. An additional annotation encoded the tail tip coordinates and the tail direction (from the tail tip point) as a value from the set of eight cardinal coordinates (N, NE, E, SE, S, SW, W, NW). The characters were stored with the coordinates of the bounding boxes of each character that emitted at least one balloon in the album. This selection aimed to retain only the main characters who had a direct influence on the story and ignore the secondary characters. Certain parts of the characters (e.g. hand, foot) were ignored in certain postures to maximize the area occupied by the character and minimize the background information in its bounding box. Also, the semantic relations between text, balloon and comic character were added.

This information was stored in a Scalable Vector Graphics (SVG) format described and available through the dataset website[4].

Metrics We evaluated the low level processing (panel, text, balloon and character extraction) in terms of object bounding boxes such as the PASCAL VOC challenge [11]. In this challenge, the detections were assigned to ground truth objects and judged to be true or false positives by measuring bounding box overlap. To be considered a correct detection, the overlap ratio a_0 between the predicted bounding box B_p and the ground truth bounding box B_{gt} must exceed 0.5 (Formula 7). In the original VOC paper, it is mentioned that "the threshold of 50% was set deliberately low to account for inaccuracies in bounding boxes in the ground truth data. For example, defining the bounding box for a highly non-convex object, e.g. a person with arms and legs spread, is somewhat subjective". The predicted objects were considered true positives TP if $a_0 > 0.5$ or false positives FP (prediction errors).

$$a_0 = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{7}$$

Detections returned by a method were assigned to ground truth objects satisfying the overlap criterion ranked by the confidence output (decreasing). Multiple detections of the same object in an image were considered false detections.

The number of TP, FP and false negative (missed elements) FN was used to compute the recall R and the precision P of each of the methods using Formula 8 and 9.

[2] https://github.com/crigaud/publication/tree/master/2015/IJDAR/

[3] http://digitalcomicmuseum.com

[4] http://ebdtheque.univ-lr.fr/database/

Table 2: Panel extraction results.

	R (%)	P (%)	F (%)
Arai [2]	58.03	75.30	65.55
Rigaud [46]	78.02	73.17	75.52
Proposed	81.24	86.55	83.81
Proposed + validation	80.69	87.03	83.74

We also computed the F-measure F for each result.

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$P = \frac{TP}{TP + FP} \tag{9}$$

The results of each contribution are detailed for each page of the eBDtheque dataset [14] in Figures 27 to 31. Note that the vertical file names on the horizontal axis correspond to the identifier of the image in the dataset [5]. The materials related to the evaluation section are available online[6].

7.1 Panel detection evaluation

We evaluated our method on the 850 panels of the eBDtheque dataset [14] "version 2014" at bounding box level.

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.

Table 2 presents the average results we obtained compared to our previous method [46] and a method from the literature [2].

The proposed panel extraction based on connected component analysis is simple to implement, and is a fast and efficient method for comics with disconnected panels (separated by a white gutter). The validation by the expert system was not significant here because the low level processing had already reached the limits of the model. Figure 27 shows the details for each image tested, which were mainly comics with gutters. Our method is not appropriate for gutterless comics (e.b. some mangas) or strip without panel borders such as those with an extra frame around several panels.

Another weakness is when panels are connected by other elements. This experiment was performed in 28 seconds for the whole dataset using one CPU at 2.5GHz (0.05s per panel on average). Note that some of the dataset images 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 $> 90\%$ of the page area was detected.

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

[6] http://ebdtheque.univ-lr.fr/references/

Table 3: Text localisation results.

	R (%)	P (%)	F (%)
Rigaud [45]	61.00	19.66	29.75
Proposed ([45]+OCR)	60.13	42.43	49.75
Proposed + validation	44.54	65.05	52.88

7.2 Text localisation and recognition evaluation

We evaluated our method for text extraction on the 4667 text lines of the eBDtheque dataset [14] "version 2014" at object bounding box level.

In our previous work [45] text extraction was evaluated on a subset of 20 pages of the eBDtheque dataset [14]. Here we applied it to the whole dataset. We used our previous method as a baseline to show an improvement in the precision of 20% when using an OCR-based filter, without a significant loss in recall. The validation by the expert system improved the precision as expected but also resulted in a drop in recall. The drop in recall 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. As in [45], this method has some difficulty coping with certain types of text that can be found in the comics e.g. graphic sounds.

We also evaluated text transcription using string edit distance [60] between a predicted text transcription given by the OCR and its corresponding transcription in the ground truth. The eBDtheque dataset is composed of English, Japanese and French texts. We evaluated the subset of English and French pages using the OCR with the corresponding training data[7]. This was performed at the text line level taking as correct the text lines that were transcribed exactly as the ground truth transcription, considering all the letters as lower case and ignoring accents (for predicted and ground truth regions). We obtained a score of 7.18% accuracy which constitute a baseline for future work on text recognition on the eBDtheque dataset [14]. We performed a more relaxed evaluation where we also considered as correct the text lines at a text edit distance equal to one, the accuracy improved to 10.46%. The distribution of the text line lengths is given in Figure 22.

Note that in Figure 22, there are more than a hundred text lines of only one letter corresponding to punctuation or single letter words such as "I" or "A"; this is a particularity of comics.

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

7.3 Balloon detection evaluation

We evaluated our method on the 1092 balloons of the eBDtheque dataset [14] "version 2014" at object bounding box level, which includes the tail. Note that our method does not require any previous processing, in contrast to [44] and it is able to detect closed balloons only. In the eBDtheque ground truth, only 84.5% of the balloons are closed and 15.5% are not. Thus we did not expect to reach 100% recall and precision.

The minimum number of children $minNbChildren$ of a balloon was set to 8, just before the first peak in the distribution of the number of letters per speech balloon in the eBDtheque dataset [14] (Figure 23). Note that in Figure 23, 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 to avoid detecting a lot of non balloon regions. Balloons with a confidence value $C_{balloon}$ lower than 10% were rejected according to the validation experiments on the eBDtheque dataset [14].

Table 4 shows the average results for the one hundred images of the dataset. We also compare to a state of the art

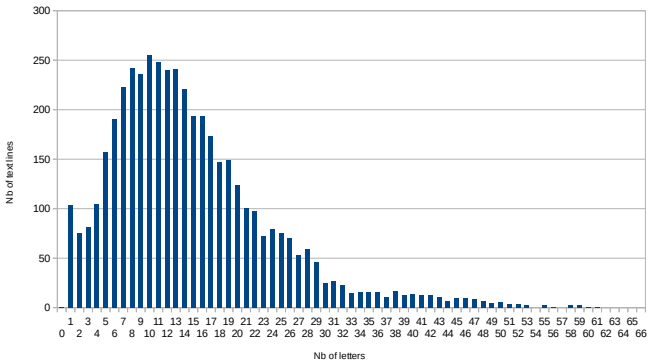


Fig. 22: Distribution of the number of letter per text lines.

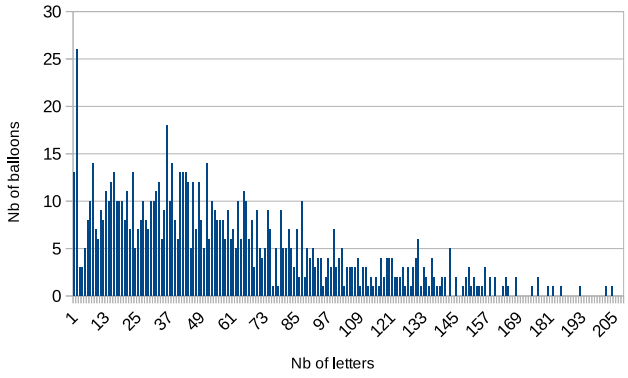


Fig. 23: Distribution of the number of letter per speech balloon.

Table 4: Balloon extraction results.

	R (%)	P (%)	F (%)
Arai [2]	6.66	10.98	8.29
Rigaud [44]	46.13	17.44	25.31
Proposed	57.90	73.84	64.91
Proposed + validation	54.79	88.76	67.75

method from the literature [2] and our previous work [44] based on the best results we had obtained for text localisation (Table 3). Our method outperforms [2] thanks to its genericity, since it can process all the image styles of the eBDtheque dataset. This was expected as [2] was specifically developed for manga comics that have certain stylistic particularities. We also surpassed our previous method [44] because it needs text lines as input which were given in our proposed text extraction method (Section 5.4). Here we clearly see the limitations of dependency between the processing. The performance of our text extractor was 49.75% (Table 3) which was used as input for balloon extraction so the balloon extraction [44] was inevitably affected. The validation of the expert system once again improve the precision but decreased the recall of the extraction while improving the overall F-measure by almost 3%. The drop in recall was due to the balloons that were correctly extracted but which contained no detectable element. Figure 30 confirms that our method works best when the balloons are closed, well segmented and with non cursive text inside. This experiment was performed in 22 minutes for the whole dataset using one CPU at 2.5GHz (2.2s per balloon on average).

7.4 Tail tip and tail direction evaluation

We evaluated the new tail extraction method on the 1092 balloons of the eBDtheque dataset [14] "version 2014".

Tail tip and tail direction are not represented by bounding boxes, therefore they can not be evaluated using equation 7. For this particular case, we defined two accuracy metrics $A_{tailTip}$, the accuracy of the detected position of the tip and $A_{tailDir}$ for the accuracy of the tail direction. The Euclidean distance d_0 between the predicted position of the tip and its ground truth was measured relative to balloon size (Formula 10). Note that we considered incorrect the predicted positions at a distance d_0 superior to the balloon size ($A_{tailTip} < 0$).

$$A_{tailTip} = 1 - \frac{d_0}{0.5 * (B_{width} + B_{height})} \tag{10}$$

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 the introduction of section 7: $A_{tailDir} = 1 - d_1/8$. 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 $C_{tail} = 0.0\%$ (confidence equal to zero percent); in this case $A_{tailTip} = A_{tailDir} = 100\%$ if there was effectively no tail to detect or $A_{tailTip} = A_{tailDir} = 0\%$. For this experiment, the local window size M of the tail direction process was set to 10% of the mean balloon size (Equation 5) in order to be invariant to the image definition. We obtained an average score of $A_{tailTip} = 96.77\%$ and $A_{tailDir} = 80.49\%$ on the eBDtheque dataset [14]. This experiment was performed in 32 seconds for the whole dataset using one CPU at 2.5GHz (0.03s per balloon on average).

Figure 29 shows the detail for each page of the eBDtheque dataset [14]. This method was very good at locating the tail tip when one and only one tail existed and also at giving a confidence value equal to zero when the balloon was completely flat (no tail). However, the algorithm was confused in case of multi-tail balloons and spiky balloons without tail because one of the peaks was detected as a tail (Figure 24 bottom). Tail direction was sensitive to the quality of tail tip position detection and the eight directions quantization. For instance, in the bottom-left part of Figure 24, an incorrect tail position was detected and thus confused the tail direction detection (detected as East instead of no direction because no tail).

7.5 Comic character detection evaluation

We evaluated the detection of comic characters on the 1550 characters of the eBDtheque dataset, given the input information from the ground truth (panel, balloon, tail position and tail direction). In the eBDtheque ground truth, only 880 (56.8%) of the character instances are speaking, thus we did not expect to reach 100% recall and precision here because our method is able to detect only speaking characters. An example of valid and rejected regions is shown on Figure 25. Cumulative results are presented in Table 5 in order to show the performance over each steps of the following process:

a) Hypothesis of ROI from Section 6.3 (A)
b) Character extraction from Section 5.5 (B)
c) Character extraction validation from Section 6.1 (C)

As shown on Figure 25, the decision criterion $a_0 > 0.5$ may seem restrictive and have to be adjusted according to the application. The numbers in Table 5 are given for the whole dataset including the characters that are not speaking. If we focus on the subset of 880 speakers, the recall, precision and F-measure were 34.49%, 32.02% and 33.21% respectively.

Table 5: Character localisation results for each step.

Step	R (%)	P (%)	F (%)
A (hypothesis)	19.28	26.91	22.47
A + B (+ extraction)	23.31	33.55	27.51
A + B + C (+ validation)	21.57	40.52	28.16

7.6 Relation retrieval evaluation

The 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 relations $STSB$ and $SBSC$ according to the metadata in the ground truth of the eBDtheque dataset [14] called *isLineOf* and *isSaidBy*, which represent 3427 and 880 relations respectively. Given the panel, balloon and character positions from the ground truth, our ontologies were consistent with 96.9% of $STSB$ assertions and 70.66% of $SBSC$. These numbers can be seen as the legitimacy value of the semantics given the relations between balloons, characters and text lines. The 3.1% of missed *isLineOf* relations came from balloons that were not compliant with our model. In the same way, of the 880 *isSaidBy* relations, that link speech balloons to speaking char-

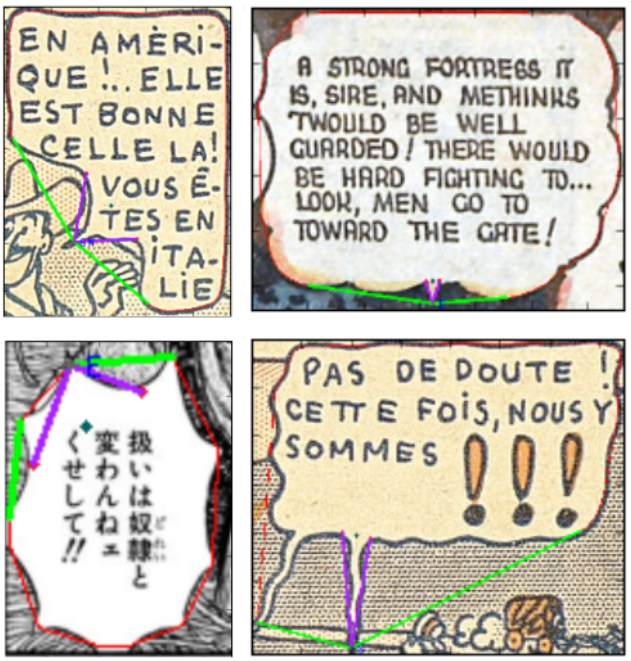


Fig. 24: Result examples of tail tip position and direction detections. Balloon contour hulls are represented in red, hull segments with the two biggest convexity defects in green and detected tails in purple. From top left to bottom right, the detect tail directions are SW, S, E and S.

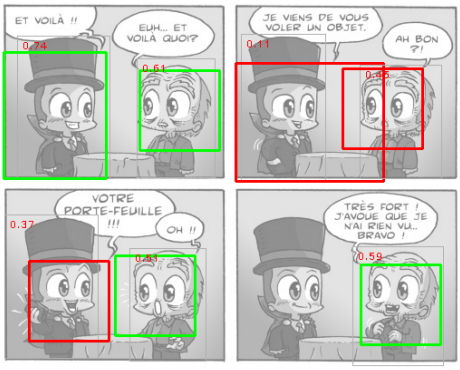


Fig. 25: Example of predicted region (hypothesis of ROI) considered as true (green) or false (red) positives according to the validation criteria $a_0 > 0.5$. Thin grey rectangles are the ground truth regions and the red numbers in the top left corner of each ground truth region is the value of the overlapping ratio a_0. Best viewed in colour.

acters, 9.5% were undetectable because they were generated from balloons outside the panel.

7.7 Framework evaluation

We evaluated our framework during the two iterations of the process loop introduced in Section 3.3 and particularly at the beginning (Step 1: hypothesis) and at the end (Step 3: inference) of each iteration. Note that using this framework some processes (C, $STSB$, $SBSC$) were related to previous processes which propagated errors and reduced their performance compared to the evaluation in sections 7.5 and 7.6. We evaluated our framework using the F-measure of panel P, balloon B, text T and character C extractions. Also, the accuracy of the two $STSB$ and $SBSC$ relations was measured. Table 6 details the number of retrieved elements among all the elements and after the two iterations of the process. The change in the performance of each processing throughout the process is presented in Figure 26 as F-measure for P, B, T, C and accuracies for $STSB$ and $SBSC$.

Process details Figure 26 shows the change of the performance of our framework after the first and second iteration of analysis over the eBDtheque dataset [14]. The performance of the first iteration was measured after the initial extraction of simple elements which were considered as hypotheses by the expert system (Figure 26a.) and after validation by the expert system (Figure 26b.). Between the first initialization and validation (first row of Figure 26), the F-measure remained stable for P and increased by 3% for the balloon B and text T. At this point, no comic book characters were discovered because their rules in the ontology are

Table 6: Result details at the end of the two iterations of the process loop. The columns "Total" correspond to the number of elements in the ground truth, "Modelled" the number en percentage of correctly modelled elements and "Detected" the number of elements effectively retrieved.

	Total	Modelled	%	Detected	%
P	850	846	99.6	586	69.22
B	1082	945	87.4	418	44.23
T	4667	3808	81.6	1941	50.97
C	1550	1457	94.9	214	14.68
$STSB$	3427	3320	96.9	1631	49.12
$SBSC$	880	621	70.7	160	25.76
TOTAL	12456	10997	88.29	4950	45.01

related to elements that were not yet discovered. Nevertheless, the links between speech text and balloon $STSB$ were inferred. The expert system applied the inference rules to automatically label the balloon with a tail that included text such as speech balloon SB and speech text ST respectively and created a $STSB$ link between each of them.

The newly inferred links were 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 could make use of rules related to characters because the speech balloons were now part of the knowledge base. Panels and speech balloons were used to create hypotheses for the location of characters and then the low level processing located them more precisely within these regions (Figure 26c.). Finally, the expert system validated the newly discovered regions and inferred the relations between speech balloons and speaking characters $SBSC$ (Figure 26d.).

Low level processing The low level processing scores (P, B, T) always increased between the hypothesis and the validation steps, which confirms the benefits of combining different levels of analysis. The best extraction performance was obtained for the panels that were usually the easiest elements to extract from a page. The lowest extraction performance was 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 link extractions) that are required to guess comic character locations.

Relations between elements The expert system was able to retrieve 49.12% of the $STSB$ and 25.76% of the $SBSC$ relations of the modelled relations. Even if these relations

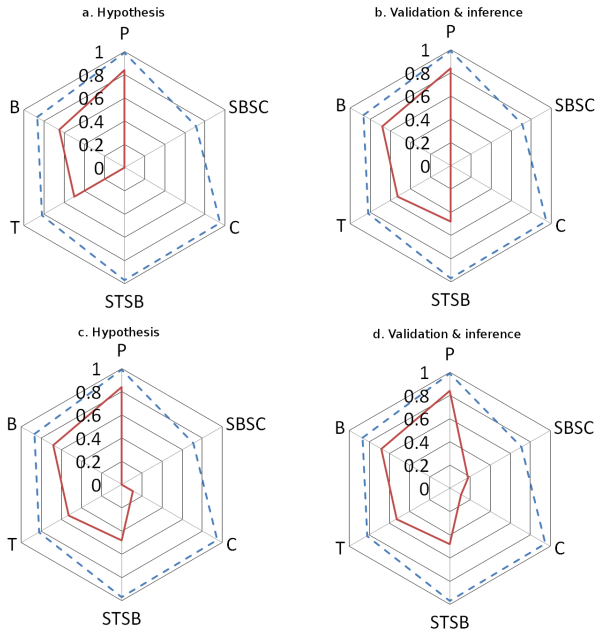


Fig. 26: Evolution of the amount of discovered information for panels P, balloons B, text lines T, comic characters C, $STSB$ and $SBSC$ extractions (solid red line). First and second row correspond to the hypothesis and evaluation steps of the first and second iterations of the process respectively. Values correspond to F-measures for P, B, T, C and accuracies for $STSB$ and $SBSC$. The dashed blue line represents the best score using our model on the data extracted from the ground truth (optimal condition). The solid red line is the performance of the framework using all the automatic extractions and elements associations presented in this paper.

were quite well modelled (96.0% and 70.7% in Table 6), their extraction remain dependent on the quality of related element extractions. It should be stressed that these numbers represent the efficiency of the last process of the whole framework pipeline (relation retrieval). Individual errors at each recognition and validation step of the pipeline are propagated to the final relation retrieval between elements. Therefore a single improvement in the detection or the validation of any kind of element would have an impact at the relation retrieval level.

General evaluation It is difficult to combine individual metrics of a different nature into a single global metric able to evaluate this framework. Table 6 gives indications about the amount of discovered information and Figure 26 provides the performance of the framework over the process loop. According to Table 6, our model was able to model 88.29% of the information of the dataset despite the diversity of the images. Moreover, 45.01% of the information were retrieved automatically using our framework (Table 6). Note that the recall and precision represented by the F-measure in Figure 26, shows good performance for simple element extraction such as P, B and T but suffers from error propagation effect for more complex ones such as C and the relations between elements ($STSB$ and $SBSC$).

8 Conclusion

This paper presents 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. We suggest improved methods for panel and balloon extraction along with a first method in the literature to locate the tail tip and the indicated direction by analysing the balloon contour. We provide a novel generic and unsupervised definition of the comic character region of interest that takes into account the spatial organisation of the rest of the elements in an image. It relies on an inference engine that interacts with two knowledge models, one for image processing and the other one for comics. In the future we plan to add more iterations of the process in order to retrieve new information such as spotting the non speaking comic characters using those already detected (speaking characters) for training. In addition, the expert system will be used to improve text extraction and recognition using system feedback in order to automatically extract open speech balloons from text locations.

Acknowledgement

The authors would like to thank Karell Bertet and Arnaud Revel for their help with the high level processing. This work was supported by a European Doctorate scholarship of the University of La Rochelle, European Regional Development Fund, the region Poitou-Charentes (France), the General Council of Charente Maritime (France), the municipality of La Rochelle (France) and the Spanish research projects TIN2011-24631, RYC-2009-05031. We are grateful to all authors and publishers of comics and manga from the eBDtheque dataset for having allowed us to show (Figure 1 to 24), use and share their works.

Compliance with Ethical Standards

The authors declare that there are no conflicts of interest. This article does not contain any studies with human or animal subjects.

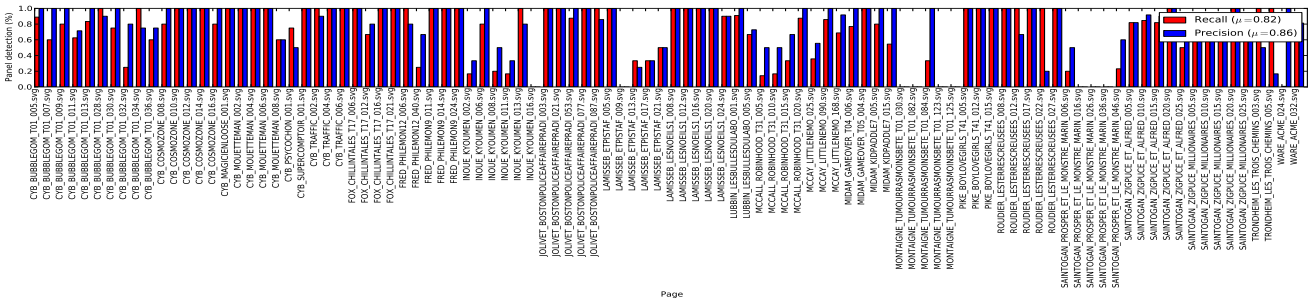


Fig. 27: Panel extraction score details for each image of the eBDtheque dataset [14] using our method in Section 7.1

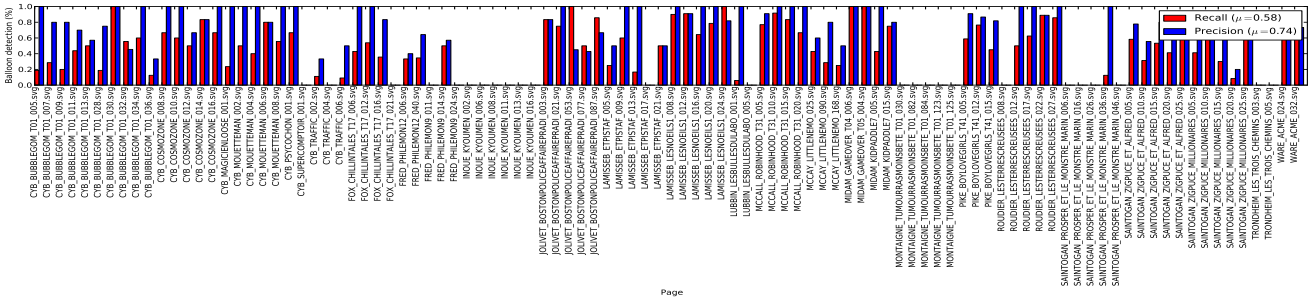


Fig. 28: Balloon extraction score details for each image of the eBDtheque dataset [14] for the evaluation in Section 7.3

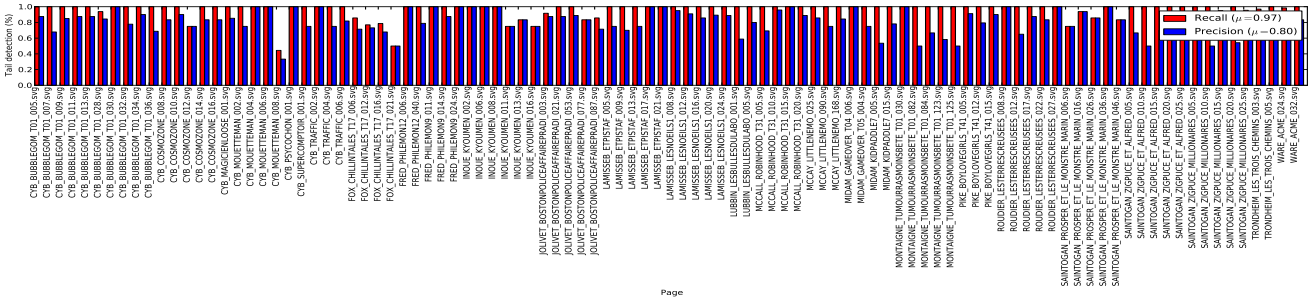


Fig. 29: Tail description score details for each image of the eBDtheque dataset [14] for the evaluation in Section 7.4.

References

1. K. Arai and H. 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. 3
2. K. Arai and H. Tolle. Method for real time text extraction of digital manga comic. *International Journal of Image Processing (IJIP)*, 4(6):669–676, 2011. 3, 8, 14, 15
3. J. B. P. Association. An introduction to publishing in japan 2012-2013, 2012. 1
4. 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. 1
5. T. Blaschke, G. J. Hay, M. Kelly, S. Lang, P. Hofmann, E. Addink, R. Q. Feitosa, F. van der Meer, H. van der Werff, F. van Coillie, and D. Tiede. Geographic object-based image analysis: Towards a new paradigm. *Journal of Photogrammetry and Remote Sensing*, 87(0):180 – 191, 2014. 2
6. M. Borodo. Multimodality, translation and comics. *Perspectives*, pages 1–20, 2014. 1
7. D. Brandon. Graphic novels and comics for the visually impaired explored in award-winning paper, 2014. 1
8. E. Di Sciascio, F. M. Donini, and M. Mongiello. Structured knowledge representation for image retrieval. *Journal of Artificial Intelligence Research*, 16(1):209–257, 2002. 2, 3, 4
9. B. Duc. *L'art de la BD - Tome 1 - Du scénario à la réalisation*. Glénat, 1982. 5
10. R. O. Duda and P. E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15:11–15, January 1972. 3
11. M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 13
12. S. Fidler, J. 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. 2
13. C. Guérin. Ontologies and spatial relations applied to comic books reading. In *PhD Symposium of Knowledge Engineering*

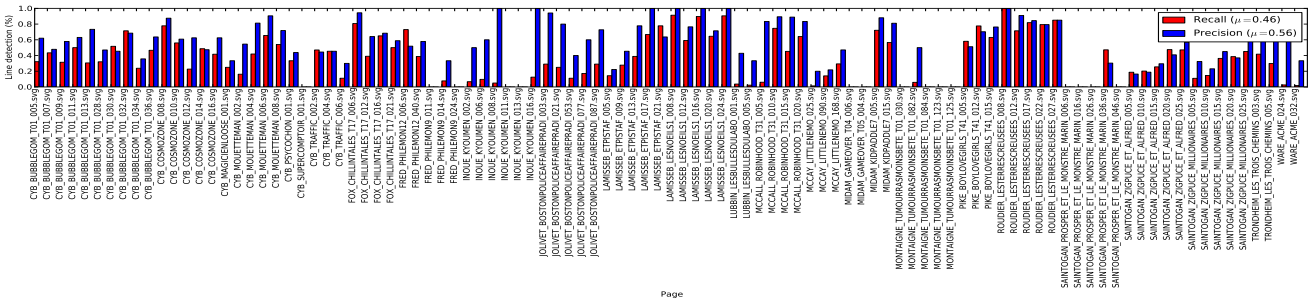


Fig. 30: Text extraction score details for each image of the eBDtheque dataset [14] for the evaluation in Section 7.2

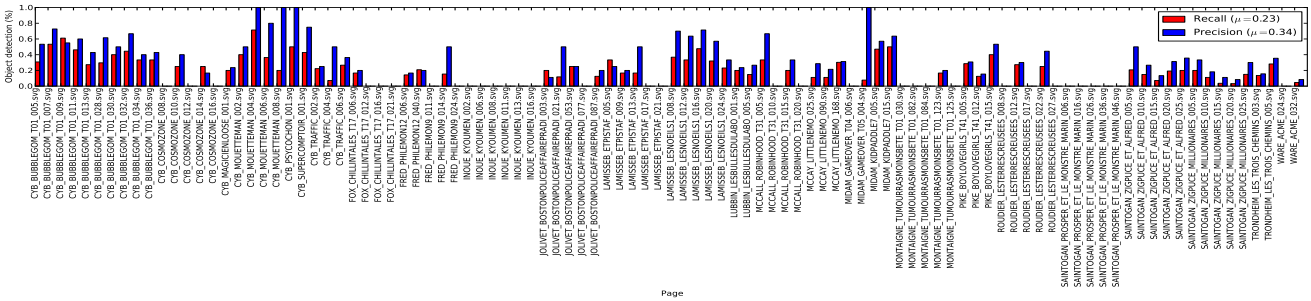


Fig. 31: Comic character extraction score details for each image of the eBDtheque dataset [14] for the evaluation in Section 7.5

and Knowledge Management (EKAW), Galway, Ireland, 2012. 3

14. C. Guérin, C. Rigaud, A. Mercier, and al. ebdtheque: a representative database of comics. In *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*, Washington DC, 2013. 5, 11, 13, 14, 15, 16, 17, 19, 20
15. V. Haarslev, K. Hidde, R. Möller, and M. Wessel. The racerpro knowledge representation and reasoning system. *Semantic Web*, 3(3):267–277, 2012. 12
16. E. Han, K. Kim, H. Yang, and K. 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. 3
17. F. Hayes-Roth, D. Waterman, and D. Lenat. *Building expert systems*. Addison-Wesley, Reading, MA, Jan 1984. 2
18. A. Hermann, S. Ferré, and M. Ducassé. Guided semantic annotation of comic panels with sewelis. In *EKAW*, volume 7603 of *Lecture Notes in Computer Science*, pages 430–433. Springer, 2012. 3
19. A. K. N. Ho, J.-C. Burie, and J.-M. Ogier. Comics page structure analysis based on automatic panel extraction. In *GREC 2011, Nineth IAPR International Workshop on Graphics Recognition*, Seoul, Korea, September, 15-16 2011. 3
20. A. K. N. Ho, J.-C. Burie, and J.-M. Ogier. Panel and Speech Balloon Extraction from Comic Books. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 424–428. Ieee, Mar. 2012. 3, 8
21. H. N. Ho, C. Rigaud, J.-C. Burie, and J.-M. 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. 3
22. B. Hu, S. Dasmahapatra, P. Lewis, and N. Shadbolt. Ontology-based medical image annotation with description logics. In *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, 2003. 2
23. C. Hudelot, J. Atif, and I. Bloch. Fuzzy spatial relation ontology for image interpretation. *Fuzzy Sets and Systems*, 159(15):1929 – 1951, 2008. 2
24. IBISWorld. Comic book publishing in the US: Market research report, 2013. 1
25. Y. In, T. Oie, M. Higuchi, S. Kawasaki, A. Koike, and H. 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. 3
26. R. Jérémy and B. 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. 1
27. F. S. Khan, M. A. Rao, J. van de Weijer, A. D. Bagdanov, M. Vanrell, and A. Lopez. Color attributes for object detection. In *Twenty-Fifth IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, 2012. 3, 11
28. J.-M. Lainé and S. Delzant. *Le lettrage des bulles*. Eyrolles, 2010. 5
29. B. Lamiroy and J.-M. Ogier. Analysis and Interpretation of Graphical Documents. In D. Doermann and K. Tombre, editors, *Handbook of Document Image Processing and Recognition*. Springer, 2014. 2
30. C. Li, A. Kowdle, A. Saxena, and T. Chen. Toward holistic scene understanding: feedback enabled cascaded classification models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1394–1408, 2012. 2
31. L. Li, Y. Wang, Z. Tang, and L. Gao. Automatic comic page segmentation based on polygon detection. *Multimedia Tools and Applications*, 69(1):171–197, 2014. 3
32. L. Li, Y. Wang, Z. Tang, X. Lu, and L. Gao. Unsupervised speech text localization in comic images. In *12th International Con-*

ference on Document Analysis and Recognition (ICDAR), pages 1190–1194, Aug 2013. 3, 8
33. S. Mao, A. Rosenfeld, and T. Kanungo. Document structure analysis algorithms: a literature survey. In T. Kanungo, E. H. B. Smith, J. Hu, and P. B. Kantor, editors, *Document Recognition and Retrieval X*, volume 5010 of *SPIE Proceedings*, pages 197–207. SPIE, 2003. 2
34. S. McCloud. *Understanding comics*. William Morrow Paperbacks, 1994. 1
35. D. L. McGuinness and F. Van Harmelen. OWL Web Ontology Language Overview. Technical report, W3C, 2004. 7
36. V. Mezaris, I. Kompatsiaris, and M. G. Strintzis. An ontology approach to object-based image retrieval. In *International Conference on Image Processing (ICIP)*, volume 2, pages 511–514, 2003. 2
37. J. 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. 2
38. N. Otsu. A threshold selection method from gray level histograms. *IEEE Trans. Systems, Man and Cybernetics*, 9:62–66, Mar. 1979. minimize inter class variance. 8
39. C. Ponsard. Enhancing the accessibility for all of digital comic books. *e-Minds*, 1(5), 2009. 1
40. C. Ponsard, R. Ramdoyal, and D. Dziamski. An ocr-enabled digital comic books viewer. In *Computers Helping People with Special Needs*, pages 471–478. Springer, 2012. 3, 10
41. G. Ratier. 2013 : l'année de la décélération — acbd.fr, 2013. 1
42. C. Rigaud, D. Karatzas, J.-C. Burie, and J.-M. Ogier. Speech balloon contour classification in comics. In *Proceedings of the 10th IAPR International Workshop on Graphics Recognition (GREC)*, pages 23–25, Bethlehem, PA, USA, 2013. 3
43. C. Rigaud, D. Karatzas, J.-C. Burie, and J.-M. Ogier. Color descriptor for content-based drawing retrieval. In *Proceedings of International Workshop on Document Analysis Systems (DAS)*, Tours, France, 2014. 3, 11
44. C. Rigaud, D. Karatzas, J. Van de Weijer, J.-C. Burie, and J.-M. Ogier. An active contour model for speech balloon detection in comics. In *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2013. 3, 8, 15
45. C. Rigaud, D. Karatzas, J. Van de Weijer, J.-C. Burie, and J.-M. Ogier. Automatic text localisation in scanned comic books. In *Proceedings of the 8th International Conference on Computer Vision Theory and Applications (VISAPP)*. SCITEPRESS Digital Library, 2013. 3, 10, 14
46. C. Rigaud, N. Tsopze, J.-C. Burie, and J.-M. Ogier. Robust frame and text extraction from comic books. In Y.-B. Kwon and J.-M. Ogier, editors, *Graphics Recognition. New Trends and Challenges*, volume 7423 of *Lecture Notes in Computer Science*, pages 129–138. Springer Berlin Heidelberg, 2013. 3, 8, 14
47. G. Robin Varnum, Christina T. *The Language of Comics: Word and Image*. Studies in Popular Culture. University Press of Mississippi, 2007. 9
48. S. Sarwar, Z. U. Qayyum, and S. 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}. 2
49. S. Singh, A. D. Cheok, G. L. Ng, and F. 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. 1
50. E. Sirin, B. Parsia, B. Cuenca Grau, A. Kalyanpur, and Y. Katz. Pellet : A Practical OWL-DL Reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):51–53, 2007. 12
51. R. Smith. An overview of the tesseract ocr engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02*, ICDAR '07, pages 629–633, Washington, DC, USA, 2007. IEEE Computer Society. 10
52. A. Stewart. A brief history of the american comic book industry, 2000. 1
53. M. Stommel, L. I. Merhej, and M. G. Müller. Segmentation-free detection of comic panels. In *Computer Vision and Graphics*, pages 633–640. Springer, 2012. 3
54. C.-Y. Su, R.-I. Chang, and J.-C. Liu. Recognizing text elements for svg comic compression and its novel applications. In *Proceedings of the 11th International Conference on Document Analysis and Recognition*, ICDAR '11, pages 1329–1333, Washington, DC, USA, 2011. IEEE Computer Society. 3
55. W. Sun and K. 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. 3, 11
56. W. Sun, K. Kise, J.-C. Burie, and J.-M. Ogier. Specific comic character detection using local feature matching. In *Proceedings of International Conference on Document Analysis and Recognition (ICDAR 2013)*, Washington, USA, 2013. 3
57. S. Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30(1):32–46, 1985. 8
58. T. Tanaka, K. Shoji, F. Toyama, and J. Miyamichi. Layout analysis of tree-structured scene frames in comic images. In *IJCAI'07*, pages 2885–2890, 2007. 3
59. E. Thomas. *Invisible Art, Invisible Planes, Invisible People*. Multicultural Comics: From Zap to Blue Beetle. University of Texas Press, 2010. 1
60. R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1):168–173, 1974. 14
61. M. Yamada, R. Budiarto, M. Endo, and S. Miyazaki. Comic image decomposition for reading comics on cellular phones. *IEICE Transactions*, 87-D(6):1370–1376, 2004. 3