# Text-independent speech balloon segmentation for comics and manga

Christophe Rigaud, Jean-Christophe Burie, Jean-Marc Ogier

# Text-independent speech balloon segmentation for comics and manga

Christophe Rigaud, Jean-Christophe Burie, Jean-Marc Ogier
Laboratoire L3i
Université de La Rochelle
Avenue Michel Crépeau 17042 La Rochelle, France
{christophe.rigaud, jean-christophe.burie, jean-marc.ogier}@univ-lr.fr

***Abstract*—Comics and manga are one of the most popular and familiar forms of graphic content over the world and play a major role in spreading country's culture. Nowadays, massive digitization and digital-born materials allow page-per-page mobile reading but we believe that other usages may be released in the near future. In this paper we focus on speech balloon segmentation which is a key issue for text/graphic association in scanned and digital-born comic book images. Speech balloons are at the interface between text and comic character, they inform the reader about speech tone and the position of its speaker. We present a generic and text-independent segmentation method based on topological and spatial organization of connected component.**

***Keywords*—graphic recognition, speech balloon, comics image analysis, manga image analysis**


(a) Rectangular speech balloon with tail and gray background (b) Wavy speech balloon with tail and yellowed background (c) Oval speech balloon with tail and clear background (d) Unusual thought balloon with tail and complex background

Fig. 1: Pixel-level speech balloon segmentation results for different balloon shape and background (red line).

## I. Introduction

The sales of digital comics is now reaching 10% of the comics market and has doubled during the last five years[1]. Such new way of reading allows new capabilities thanks to the richness of the drawings and the recent development of mobile platform reading tools. Apart from layout re-flowing (panel re-arrangement) according to screen size, there are few work exploring other ways of reading.

In this paper, we are interested in pixel-level balloon segmentation in order to retrieve position and shape of the speech balloons. Both information are key issue for balloon classification [1] and comic character association [2]. This last information is not implicitly put by the cartoonist into the drawing but understood by the reader according to the position of the elements in the images. Speech balloons are placed in a way that helps the reader to associate them with comic characters and to follow the story easily. The classical comics design is, speech text placed inside balloons, balloons are contained by panels with the concerned comic characters (emitter of the balloon) close to it. Panel, balloon and comic character positions are the three information required to associate speech balloons and comic characters and start to towards comics understanding. Panel extraction is the easiest task in comics image analysis and several studies exceed 80% recall and precision [3], [4]. Balloon extraction attracted little attention even-though it is an helpful information for text extraction and essential for text/graphics association. Comic character extraction is at its early stage and the information of speech balloon positions together with their tail can help a lot for generic methods designed for such complex graphics extraction [2]. A pixel-level speech balloon extraction appear to be important for further processing such as balloon contour and tail analysis Fig. 1.

We propose a text-independent approach and parameter-free speech balloon extraction appropriate to any closed balloon which are the most common type of balloon. We base our approach on the observation that speech balloons almost always contain vertically or horizontally aligned elements (text property).

Balloons or bubbles are key elements in comics, they contain most of the textual information and go pairwise with comic characters (speakers). Few works about balloon extraction have been done until now and mainly closed speech balloons have been studied (balloons with a fully connected outline). Arai [5] proposed a white blob detection method based on connected-component detection with four filtering rules related to manga image analysis. The rules are based on blob size, white pixel occurrence, presence of vertical spaces and width to length ratio. Another connected-component approach proposed by Ho [6] uses HSV color space to make a first selection of bright blobs and then consider as balloons the blobs with a ratio between the text area and the blob bounding box higher than 60%.

Our group developed a method to extract open balloons (balloons with partial outline) by inflating a contour model around text regions [7]. This approach requires text positions as input which are used to initialize the active contour around text areas.

Section II presents the proposed speech balloon segmen-

[1]Milton Griepp's White Paper, ICv2 Conference 2014

tation method. Section III the experiments we performed. Finally, Section IV and V discuss and conclude this work respectively.

## II. Speech balloon extraction

From the three approaches reviewed in Section I, the first approach have been developed especially for manga and therefore has several weaknesses for other types of comics. First, the extraction of the connected-components (CC) requires a binary image which is obtained by using a global threshold. This limits its application to images with clear background color (close to white). Second, balloon candidates selection is performed using several heuristics which are not validated and specific to manga. The method proposed by Ho [6] can be very efficient for a particular comics type but the set of parameter makes it not adaptive to all styles of comics and manga (e.g. heuristic of minimum percentage of text inside balloons). Our previous method [7] requires text positions as input which is a strong constraint but has the advantage to retrieve open-balloons as well.

We propose to overcome these limitations by using a local and adaptive threshold selection method in order to binarize the gray-level version of the image and extract the connected-components. The advantage of using a local and adaptive threshold selection method is to limit original strokes to be broken after the binarization 2. After having extracted all the CC, we select only the ones with a particular content topology and alignment, independently from size and script (written signs) and compute an overall confidence value that is used for the final decision.

### A. Adaptive threshold selection

During the comics or manga creation process, balloon outlines are first drawn using a black stroke and then they are filled with text [8].

We propose to rely on these two information as they are characteristics of speech balloons.

The outlines are intentionally created in a continuous way by the artist whether they are straight or curved (single stroke). Sometimes, they appear to be degraded when reaching the final reader (e.g. image digitization, compression). A perfect outline segmentation and connected-component extraction would result in one single connected-component per outline segment Fig. 2.

However, the regions with complex background complicates this step. There are several adaptive threshold selection method in the literature [10]. However, speech balloon being a highly contrasted region it is not very complex to separate background and content (foreground) in this region. The main difficulty is the size of the region ($blockSize$) which is used to determine if a pixel belongs to the background or foreground. We define the threshold value $T = (x, y)$ as the mean of the $blockSize * blockSize$ neighborhood of $(x, y)$. The corresponding pixel at position $(x, y)$ is considered as part of the foreground if its gray-value is above $T$ or background else. The $blockSize$ is defined as a square of area 1.3% of the image area according to the validation on the eBDtheque dataset [11].

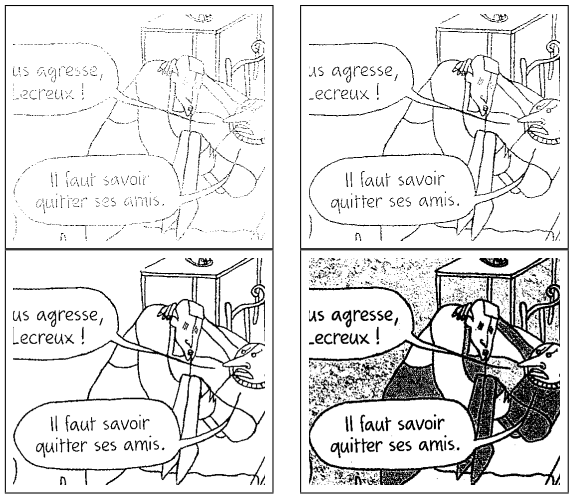
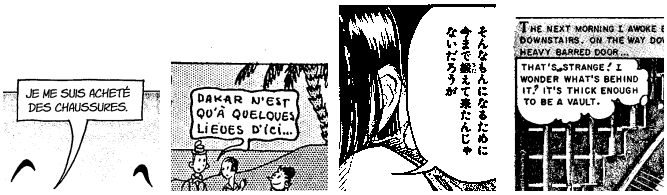



Fig. 2: Binarization results of a 8 bits gray image at different threshold levels from the lower (top-left) to the higher (bottom-right) with threshold = 50, 100, 150, 200. We observe that the black strokes are broken at a low threshold and the background starts to appear as salt and paper noise due to paper texture at high threshold value. The best binarization corresponds to threshold = 150 here. Source: [9].




(a) Recent French comics (A4 page digitized at 300DPI)
(b) Old French comics (A4 page digitized at 300DPI)
(c) Recent manga (digital-born with JPEG compression 150DPI)
(d) Golden age American comics (A4 page digitized at 300DPI)

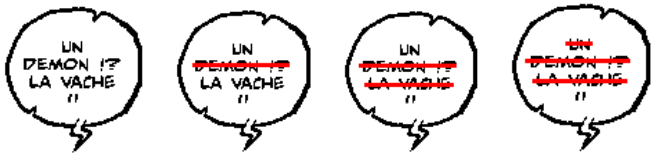
Fig. 3: Binarization results of comics and manga from different nature and definition.

Figure 3 shows binarization results of comics and manga from different nature and definition using this approach.

After having applied the selected threshold to the gray-scale image, we obtain a binary image from which we extract the CC (balloon candidates). In the next subsection, each CC content will be analyzed to determine if it contains text-like information (CC which are aligned).

### B. Candidate balloon content analysis

We propose to analyze the content organization of each candidate CC in order to determine if it contains speech text-like information (assuming speech balloons contain speech text). Speech text has several characteristics, some from the text domain and other from comics domain. Text information is characterized by, independently from language, aligned and equally separated glyphs with noticeable contrast to their background, with constant stroke width (thickness), similar color and sizes [12]. When text is used for speech balloon




(a) 0/18 aligned, start (b) 7/18 aligned, 1 line, continue (c) 14/18 aligned, 2 lines, continue (d) 16/18 aligned, 3 lines, stop

Fig. 4: Children CC alignment scanning process. The process stops automatically when there is less remaining CC than discovered lines (in the most right figure three lines have been discovered and only two children remains).

purpose, it is most of the time centered whatever the language or type of comics/manga (comics domain). The difficulty for comics analysis is the huge amount of graphics that also contain aligned elements (e.g. roofing tile, grass, hairs and eyes) which confuse text/graphics separation.

In the following approach, we combine inside balloon CC alignment and centering to compute a confidence value for each balloon candidate. The confidence value is used for the final balloon/non-balloon candidate decision (Section III).

Thereafter, we call each balloon candidate CC "parent" and the set of inside-balloon CCs "children".

*1) Alignment:* The children are supposed to be horizontally or vertically aligned according to the language (e.g. vertical for Japanese, horizontal for English or French). This is a characteristic of speech text in comics and also for text in general. We propose to "scan" (horizontally or vertically) each candidate content (children region) and compute the percentage of children which are aligned ($cAlign$). The scanning orientation (vertical or horizontal) is defined manually here but it could be replaced by an automatic detection based on parent elongation measurement (speech balloon containing vertical text are usually higher than large contrary to balloons containing horizontal text). We compute the percentage of aligned children in a specific order, from the longest to the shortest line of children in order to find the longest lines first (the most discriminative). The process stops automatically when there is less remaining non-aligned children than the number of lines found. See Fig. 4.

*2) Centering:* Apart from the children alignment, we also consider the global children block centering in the parent which is also intrinsic to the comics design. We use the centering as a clue about the probability of a parent CC to be a balloon, without considering the nature of the children (no previous text extraction required). Centering is calculated by comparing the gaps between the top, right, bottom and left sides of the bloc formed by the children bounding box and the parent bounding box (Fig. 5).

Each gap is measured using the Euclidean distance and two percentages are computed for vertical $vCenter$ and horizontal $hCenter$ centering respectively (Equations 1 and 2). The maximal centering values are obtained when $d_1 = d_3$ and $d_2 = d_4$ ($hCenter = vCenter = 100\%$).

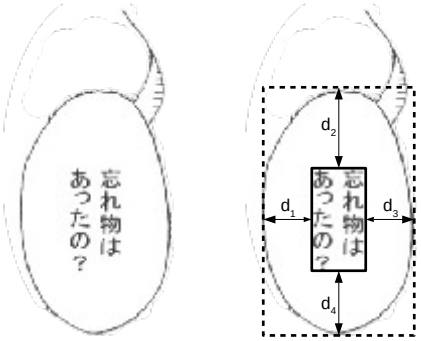



Fig. 5: Horizontal and vertical centering measurement between a parent CC and its children.

$$hCenter = 1 - \frac{|d_1 - d_3|}{d_1 + d_3} \quad (1)$$

$$vCenter = 1 - \frac{|d_2 - d_4|}{d_2 + d_4} \quad (2)$$

Both differences $hCenter$ and $vCenter$ are normalized values between zero and one, as a percentage of distance for lateral and vertical graphics.

*3) Confidence value:* The global confidence value $C$ is computed for each parent CC, combining the three indications (inter-child alignment, horizontal and vertical children bloc centering) according to Formula 3. Note that alignment ($cAlign$) and overall centering ($hCenter + vCenter$) count for half of the final confidence value respectively.

$$C = \frac{1}{2} * cAlign + \frac{1}{4} * hCenter + \frac{1}{4} * vCenter \quad (3)$$

## III. Experiments

In this section we evaluate the proposed method of speech balloon segmentation using two datasets and compare our results to other approaches from the literature.

### A. Datasets

We evaluate the proposed method using the public eBDtheque [11]. This dataset was designed to be as representative as possible of the comics diversity, including few pages of several album types. It is composed by one hundred images which are composed by 850 panels, 1550 comics characters, 1092 balloons (84.5% are closed) and 4691 text lines. It contains images scanned from French comic books (46%), French webcomics (37%) with various formats and definitions, public domain American comics (11%) and unpublished artwork of manga (6%). In addition to the diversity of styles, formats and definitions, there are also differences in design and printing techniques since 29% of the images were published before 1953 and 71% after 2000. These differences highlights the wide range of applications of the presented method.

(a) Grass aligned as text in an balloon-like region

(b) Face containing aligned and centered elements

Fig. 6: Examples of failure cases of the proposed approach.

### B. Performance evaluation

We evaluate the performance of the proposed method with common evaluation measures of recall, precision and $F_1$ score at pixel-level. Recall ($R$) is the number of correctly labeled pixels divided by the number of pixel to retrieve from the ground truth (pixel from speech balloon regions). Precision ($P$) is the number of correctly labeled pixels divided by the number of labeled pixels. Result details are given Table I.

TABLE I: Speech balloon segmentation performance in percent.

| Method | $R$ | $P$ | $F_1$ |
|---|---|---|---|
| Arai [5] | 18.70 | 23.14 | 20.69 |
| Ho [6] | 14.78 | 32.37 | 20.30 |
| Rigaud [7] | **69.81** | 32.83 | 44.66 |
| Proposed | 62.92 | **62.27** | **63.59** |

The proposed approach as 6.89% drop in recall compared to the best method from the literature mainly because it is not designed to segment open balloons which represent 15.5% of the dataset. However, our approach has the advantage do not require text position contrary to [7]. Both precision and $F_1$ measure are surpassed by 29.44% and 18.93% respectively compared to the best result from the literature which demonstrates the robustness of the proposed approach. Correct segmentation results are shown Fig. 1 and failure cases in Fig. 6.

## IV. Discussion

The proposed method uses a simple adaptive thresholding approach which is efficient for the studied scope because speech balloon regions has a simple background easy to binarize when taken apart from the rest of the image (local thresholding). Speech balloon candidate selection is based on content analysis which sometimes ends up with false positives because other graphics have a similar-to-text organization. Open balloons can be included to this approach by analyzing page background content (contains open balloon text). Note that the presented approach segments balloon background regions (white regions) but some processing require balloon outlines (e.g. tail direction retrieval). In such situation the proposed results have to be post-processed in order to find the external edge of the outline stroke.

## V. Conclusion

This paper presents a speech balloon segmentation approach toward comics and manga text/graphics association. The proposed method combines topological and spatial position relationship in order to segment speech balloon at pixel-level. The proposed approach has been tested over most of the comics type with promising performance. In the future we plan to add more scanned and digital-born manga into the dataset. Also, we would like to include open balloons segmentation in the method, based on text-like element organization in the image background.


## Acknowledgment

This work was supported by the University of La Rochelle (France) and the town of La Rochelle.We are grateful to all authors and publishers of comics and manga images from eBDtheque datasets for allowing us to use their works.


## References


[1] C. Rigaud, D. Karatzas, J.-C. Burie, and J.-M. Ogier, "Adaptive contour classification of comics speech balloons," in *Graphics Recognition. Current Trends and Challenges*, ser. Lecture Notes in Computer Science, B. Lamiroy and J.-M. Ogier, Eds. Springer Berlin Heidelberg, 2014, vol. 8746, pp. 53–62. [Online]. Available: http://dx.doi.org/10.1007/978-3-662-44854-0_5

[2] C. Rigaud, N. Le Thanh, J.-C. Burie, J.-M. Ogier, M. Iwata, E. Imazu, and K. Koichi, "Speech balloon and speaker association for comics and manga understanding," in *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR). To be published.* IEEE, 2015.

[3] M. Stommel, L. I. Merhej, and M. G. Müller, "Segmentation-free detection of comic panels," in *Computer Vision and Graphics*. Springer, 2012, pp. 633–640.

[4] L. Li, Y. Wang, C. Y. Suen, Z. Tang, and D. Liu, "A tree conditional random field model for panel detection in comic images," *Pattern Recognition*, vol. 48, no. 7, pp. 2129 – 2140, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320315000308

[5] K. Arai and H. Tolle, "Method for real time text extraction of digital manga comic," *International Journal of Image Processing (IJIP)*, vol. 4, no. 6, pp. 669–676, 2011.

[6] A. K. N. Ho, J.-C. Burie, and J.-M. Ogier, "Panel and Speech Balloon Extraction from Comic Books," *2012 10th IAPR International Workshop on Document Analysis Systems*, pp. 424–428, Mar. 2012.

[7] 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.

[8] Cyb, *Making Comics: Storytelling Secrets of Comics, Manga and Graphic Novels*. William Morrow Paperbacks, 2006, pp. 128–153.

[9] N. Roudier, *Les terres creusées*, N. Roudier, Ed. Actes Sud, 2011, vol. Acte sur BD.

[10] B. Lamiroy and J.-M. Ogier, "Analysis and interpretation of graphical documents," in *Handbook of Document Image Processing and Recognition*, D. Doermann and K. Tombre, Eds. Springer, 2014.

[11] 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.

[12] L. G. i Bigorda and D. Karatzas, "A fast hierarchical method for multi-script and arbitrary oriented scene text extraction," *CoRR*, vol. abs/1407.7504, 2014. [Online]. Available: http://arxiv.org/abs/1407.7504