# Speech balloon and speaker association for comics and manga understanding

Christophe Rigaud, Nam Le Thanh, J.-C. Burie, J.-M. Ogier
Laboratoire L3i
Université de La Rochelle
Avenue Michel Crépeau, 17000 La Rochelle, France
{christophe.rigaud, tle05, jcburie, jmogier}@univ-lr.fr

Motoi Iwata, Eiki Imazu, Koichi Kise
Graduate School of Engineering
Osaka Prefecture University
1-1 Gakuen-cho, Nakaku, Sakai, Osaka 599-8531, Japan
{iwata, imazu_e, kise}@cs.osakafu-u.ac.jp

***Abstract*—Comics and manga are one of the most important forms of publication and play a major role in spreading culture all over the world. In this paper we focus on balloons and their association to comic characters or more generally text and graphic links retrieval. This information is not directly encoded in the image, whether scanned or digital-born, it has to be understood according to other information present in the image. Such high level information allows new browsing experience and story understanding (e.g. dialog analysis, situation retrieval). We propose a speech balloon and comic character association method able to retrieve which character is emitting which speech balloon. The proposed method is based on geometric graph analysis and anchor point selection. This work has been evaluated over various comic book styles from the eBDtheque dataset and also a volume of the Kingdom manga series.**

## I. Introduction

The amount of digital manga read per day is growing very fast. Such a new way of reading allows new capabilities thanks to the richness of manga drawing and the recent progress of mobile platform reading tools. Apart from layout re-flowing (panel re-arrangement) according to screen size, there are little work exploring other ways of reading. New reading ways using new technologies requires a high level of understanding of the image content but almost no research about such level of understanding has been done before. An important and basic step toward the understanding is to associate text to graphics. In the case of comics, this means the association of text in speech balloons with comic characters as its speaker (Fig. 1). This association is not implicitly put by the manga artist into the drawing but understood by the reader according to the position of the elements in the page. Speech balloons are placed in a way that helps the reader to associate them to comic characters and follow the story fluently. The classical scenario is as follows: (i) speech text is inside speech balloons (ii) speech balloons are nearby their speakers (iii) related balloons and comic characters are (most of the time) part of the same panel. Panels, balloons and comic characters positions are the three information required to associate speech balloons and comic characters. Note that sometimes, balloons have a pointer directed towards the comic characters, more precisely its mouth, it is called "the tail".

Low level content extraction has been investigated since about ten years for comics but few recent work concern comics understanding. Recent researches are trying to reach a higher level of description by including semantic aspects. The proposed work is to bridge the gap between low and high levels by taking into account the "semantic" relationships between two physical objects, which are balloons and characters. We believe that content extraction and association methods combined with natural language processing (NLP) [1] in a holistic understanding context can benefit to comics understanding. Such combination can be useful for reconstructing timestamps from panel order, generate script and dialogs from text transcription and balloon order, etc.

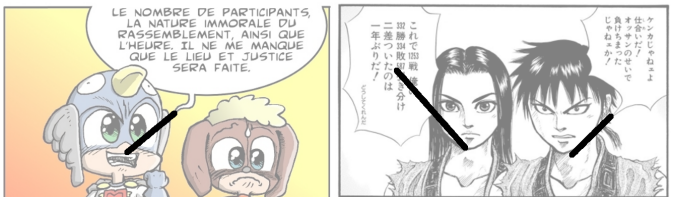



Fig. 1: Relations between speech balloons and comic characters. The relations are represented by three black lines connecting balloon centers (or tails if any) and speaker centers. Image credits: acknowledgment section.

In the next section we review content extraction and association methods. Section III details our association approach and Section IV the experiments we performed to validate the proposed method. Finally, Section V and VI discuss and conclude this work respectively.

## II. Related work

### A. Panel extraction

Several techniques have been developed to automatically extract panels, assuming that panels are small enough elements to be comfortably read on mobile devices. Several approaches are based on white line cutting with Hough transform [2], recursive X-Y cut [3] or density gradient [4]. These methods do not consider empty area and implicit panels (outline-free) [5]. These issues have been corrected by connected-component approaches but the latter are sensitive to regions that sometimes connect several panels [6]. This over-connection issue has been considered by Khoi [7] using morphological analysis and region growing method to remove such connecting elements but sometimes, it also creates new holes in the panel outline. Recently, Pang [8] introduced a new algorithm able to find splitting lines. Other methods have shown interesting

results for manga and European comics considering different background colors. They are based on watershed [9], line segmentation and polygon detection [10], region of interest detection such as corners and line segments [11].

### B. Balloon and tail extractions

Balloons are key elements in comics. They contain most of the text and go pairwise with comic characters. Little work concerns balloon extraction and mainly closed speech balloons have been studied (balloons with a fully connected outline). Arai [12] proposed a white blob detection method based on connected-component detection with four filtering rules related to manga image analysis. Another connected-component approach proposed by Ho [13] uses HSV color space to make a first selection of bright blobs and then consider as candidate the blobs above a certain ratio between the text area and the blob bounding box. Our group developed a method to extract open balloons (balloons with partially drawn outline) by inflating an active contour model around text regions [14]. Recently, we published a first approach to detect tail position and direction on the balloon contour. It is calculated by analyzing the variations between the balloon contour and a smoothed version of the balloon contour [15].

### C. Comic character extraction

Comic characters are important elements in comics. They have various features on their shapes, textures and so on. The features obtained from a same character are also various because of its various posture and facial expression. Therefore, comic character extraction is a challenging problem. Few work about comic character extraction have been published until now. Sun [16] proposed a similar partial copy recognition method for line drawings using concentric multi-region histograms of oriented gradients for copyright protection. In Sun's method, face regions obtained by Viola-Jones algorithm [17] are used as region of interest. Recently, we improved Sun's method to fit character retrieval for manga. It is more robust against various postures and facial expressions using a few labeled data [18]. We also proposed a generic and unsupervised character extraction approach from contextual information which estimates character regions of interest from the positions of balloons inside panels [15].

## III. Speech balloon and comic character association

Our intention for speech balloon and comic character association (or link) retrieval is to propose a generic approach that can deal with as many types of comics as possible. The subsequent difficulty caused by the consideration of several types of comics is the heterogeneity of the image content. The links being related to balloons and comic characters which are also related to panels, make the task particularly challenging. The proposed method is intended to be robust against all the variations generated by those related elements.

Panels represent snapshots in the story, they are often surrounded by frames and define the layout of the page. However, panels are sometimes implicit (not drawn) and methods from the literature have difficulties to localize them [6], [8]. Similarly to panels, the balloons do not always have a complete

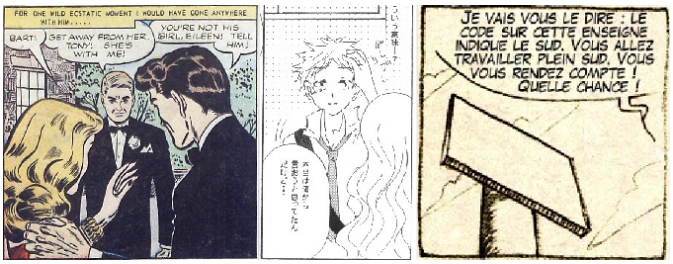
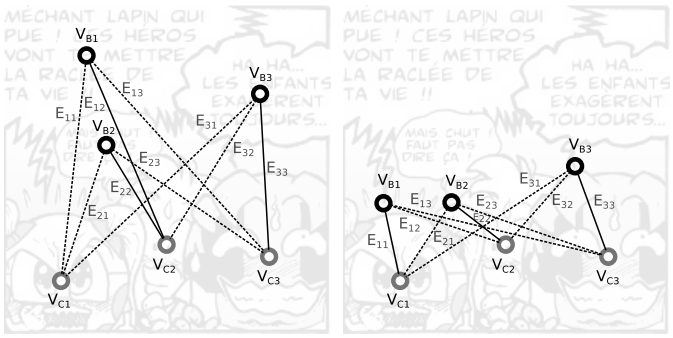



Fig. 2: Different levels of difficulty for speech balloon and comic character association. From left to right and easy to hard, a panel containing three balloons with tail and three comic characters, a panel containing one balloon without tail and two characters, a panel containing one balloon with tail and no comic character. Image credits: acknowledgment section.

outline, sometimes there is only speech text nearby comic characters [14]. The balloon tail is also not always present on the balloon boundary, it relies on artist's choices. About speaking characters, they may not appear in the image (or panel) when saying something but the reader still can follow the story by interpreting speech text into their context. Finally, the links between speech balloons and speaking characters are not explicitly drawn, they have to be interpreted by understanding the semantic of the image (Fig. 2). All those reasons are challenges for speech balloon and speaker association. We propose a method that is able to retrieve the semantic links between speech balloons and speaking characters using geometric graph analysis and anchor point selection. Note that the proposed approach assumes that panel, balloon and comic character positions have already been extracted (prerequisites).

### A. Association method

A close analysis of comics and manga images led us to the following conclusions. In order to be easily associated by the reader, balloons and comic characters are usually close to each other inside panels and sometimes, a tail directed towards the character's face or mouth is added to the balloon. This can be formulated as an optimization problem where we search for the best pairs (2-tuples) of speech balloons and comic characters corresponding to associations in the story. The main optimization criterion is the Euclidean distance between each entry of the 2-tuples but other criterion like the angle of the tail can be combined as well. Here we use only the Euclidean distance to simplify the problem and formalize it using geometric graph theory [19]. The formalism of geometric graph is appropriate for our study because it defines graphs in a Euclidean plane. Vertices (nodes) are points in general positions (balloons and characters) and edges are straight-lines segments (associations). Note that by using such representation, the angles between edges correspond to angles between elements in the image. We build a geometric graph $G = (V, E)$ where $V$ denotes the set of vertices and $E$ the set of weighted edges in each plane. In our case, the set of vertices $V$ is composed by two subsets corresponding to spatial positions of balloons $V_B \subseteq V$ and comic characters $V_C \subseteq V$ ($V = V_B \cup V_C$). It can be assimilated as a bipartite graph. We define the set of weighted edges $E$ corresponding to all the possible associations between vertices $V_B$ and $V_C$. The weight




(a) Anchor points of level 2 (b) Anchor points of level 3

Fig. 3: Geometric graphs computed from a panel with two different sets of anchor points (Table I). Balloon and comic character vertices are represented as black and gray circles respectively. Optimal associations are represented by solid straight-lines while other associations are dashed. In this example, comic characters are composed of faces only, therefore there is not difference between anchor points of level 2 and 3.

TABLE I: Examples of anchor point selections for speech balloons and comic characters, ordered by level of precision.

| Level | Speech balloon $V_B$ | Comic character $V_C$ |
|---|---|---|
| 1 | Bounding box center | Bounding box center |
| 2 | Balloon centroid | Character centroid |
| 3 | Tail tip position | Face center |
| 4 | Tail position and direction | Mouth center |

of each edge is the Euclidean distance (straight-line) between the first and the second entry of the pair (Fig. 3). In order to retrieve the associations, we associate each vertex $V_{Bi}$ with the vertex $V_{Cj}$ having the minimal association cost (weight of edge $E_{ij}$). Note that we do not consider the other way around ($V_{Cj} \rightarrow V_{Bi}$) because not all comic characters are speaking. We call the optimal association $L^*$ (Formula 1).

$$L^*(V_{Bi}, V_{Cj}) = \text{ArgMin}(E_{ij}) \tag{1}$$

For instance in Fig. 3a, the three optimal pairs that have been found are $(V_{B1}, V_{C2})$, $(V_{B2}, V_{C2})$ and $(V_{B3}, V_{C3})$ because $\text{ArgMin}(E_{ij}) = E_{12}$, $E_{22}$ and $E_{33}$ respectively. Here, the second association $(V_{B1}, V_{C2})$ is not correct because $V_{B2}$ should be associated to $V_{C1}$ according to the image. This is corrected in Fig. 3b by using a different set of anchor points.

### B. Anchor point selection method

There are several ways to define the anchors of speech balloons and comic characters. Table I gives a non-exhaustive list of anchor choices ordered by level of precision from zero to four, four being the most precise level. In Table I, we distinguish bounding box and centroid (center of mass) anchor points because their position may vary significantly according to the shape (e.g. balloon with a big tail). The anchor points are selected from the highest level available according to the information given as input to the proposed method. For instance, if the balloon extractor is able to provide balloon and tail tip positions (if any) then we will select tail tip position as anchor point. Another example, if a comic character extractor provides body and face positions then we will use face centers as anchor points.

## IV. Experiments

In this section we evaluate the proposed method of speech balloon and comic character association using two datasets. Because the proposed method relies on the quality of previous extractions (prerequisites), we evaluate our approach using two scenarios. At first, we extract prerequisites using methods from the literature. Then, we use prerequisites from ground truth in order to highlight the performance of the proposed method using error-free prerequisites.

### A. Datasets

We evaluate the proposed method using two different datasets. The first one is the public dataset eBDtheque [20], this dataset was designed to be as representative as possible of the comics diversity, including few pages of several album types. The second dataset is non public dataset composed by digital images of the first volume of the Kingdom manga (Japanese manga series written and illustrated by Yasuhisa Hara). The main difference between the two datasets is that speech balloons from eBDtheque dataset almost always have a tail contrary to the Kingdom dataset. This difference allows us to show the robustness of the presented method.

*1) eBDtheque dataset:* It is composed by one hundred images which are composed by 850 panels, 1550 comics characters, 1092 balloons 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. From this dataset, we considered only the subset of speech balloons being associated with at least one comic character, which represents 876 balloons (and associations) in total and 97% have a tail. Note that the links in the ground truth files are stored as a metadata named *linkedToCharacter* which is associated to each speech balloon region and refers to the identifier of the corresponding comic character.

*2) Kingdom dataset:* Kingdom is a famous Japanese manga drawn by Yasuhisa Hara published by Shueisha Inc. since 2006. It won the Grand Prize of Tezuka Osamu Cultural Prize in 2013. The setting of Kingdom is old Chinese history in B.C. 3rd century. We built the ground truth of the first volume of the Kingdom manga in order to evaluate the proposed method, especially its robustness to speech balloons having a small or no tail (more frequent in manga). This volume is composed of 8 chapters and consist in 216 pages in total, each digital image contains a double pages (108 images). This dataset is composed by 1159 balloons (80.33% have a tail) and associations between speech balloons and comic characters.

### B. Performance evaluation

The predicted links are considered as true if the associations between the corresponding balloons and characters exist in the ground truth. The accuracy of the proposed method is calculated from the number of retrieved elements divided by the total number of elements. As mentioned in the introduction of this section, we performed two levels of evaluation in order to highlight error sources. The first evaluation scenario consists in extracting the prerequisites (panel, balloon and comic character positions) using automatic extractors from the literature and then compute the associations using the proposed method on top of it. In this first scenario, the measured error will be a combination of errors from the proposed method and other element extractions (e.g. missed speech balloons, missed comic characters, over-segmentation of panels). In order to measure the proportion of errors which is only related to the proposed method, we perform a second evaluation. This second scenario consists in loading prerequisites from the ground truth and then only calculates the associations using the proposed method. This scenario allows us to evaluate the performance of the proposed method apart from other sources of errors because the prerequisites are error-free (from ground truth).

For each scenario, we compared two anchor point selection methods as discussed in Section III. The anchor points are computed without using (*Method A*) and using (*Method B*) the information about tail position. More precisely, for *Method A* the distance is calculated between balloon centroids and comic character centroids. For *Method B*, the distance is calculated between tail tip positions and comic character centroids. Note that even when considering the tail information (*Method B*), if the tail is missing then the results for the concerned speech balloon will be equivalent to *Method A*.

*1) Prerequisites from automatic extractions:* In this first experiment scenario, we use prerequisites from automatic content extractors in the literature. Then we associate balloons and comic characters and compare the results on the two datasets. We use our previous methods for the extraction of panels [6], balloons [14] and comic characters [18]. Note that we use another recent work from our team [15] only for extracting comic character from eBDtheque dataset considering the tail (*Method B*). This alternative is based on regions of interest computation from tail positions and directions. It is more powerful than the face-based algorithm [18] when tail information is available.

The two first rows of Table II show the results on the two datasets for *Method A* and *Method B* from automatically extracted prerequisites. It should be stressed that these numbers represent the efficiency of the association method after all prerequisite element extractions (depends on their performances). Using *Method A* (not considering tail positions), we were able to retrieve 4.33% of the associations from the eBDtheque dataset and 18.60% from the Kingdom dataset. The difference is due to the fact that the method used for comic character extraction gives poor results on non-manga images because it based on face detection trained with manga faces (more stable than other type of comic character faces). In *Method B*, the additional information of tail position is beneficial to retrieve more associations, especially for the eBDtheque dataset, but the prerequisite errors still impact a lot of results (missed balloons or missed comic characters). In the next evaluation scenario, we disregard prerequisite errors in order to highlight the part of error related to the proposed method only.

TABLE II: Accuracy of the proposed method using different prerequisites and anchor selection methods.

| Prerequisites | Anchors | eBDtheque | Kingdom |
|---|---|---|---|
| Extracted | *Method A* | 4.33% | 18.60% |
| Extracted | *Method B* | 18.01% | 19.41% |
| Loaded | *Method A* | 78.58% | 76.35% |
| Loaded | *Method B* | 93.32% | 87.74% |

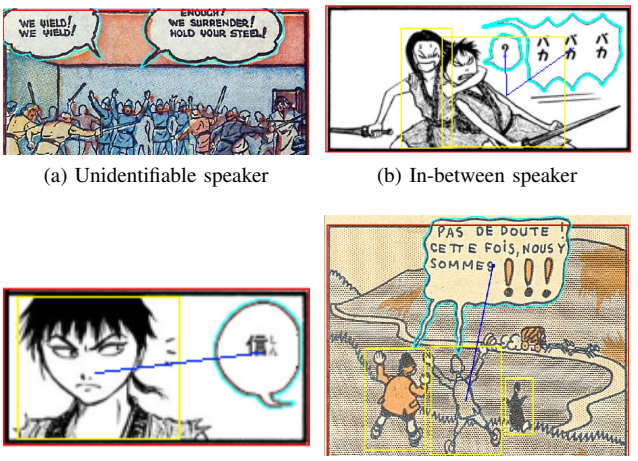
*2) Prerequisites from ground truth:* Given panel, balloon and comic character regions from the ground truth (error-free), we evaluated the performance of the proposed method at retrieving the links between speech balloons and comic characters for both datasets. Results are given in the third and fourth row of Table II. Third row of Table II (anchors from balloon and comic character centroids), shows similar results for both datasets. This means that our method is robust to comic type variations (eBDtheque is mainly composed by European comics and Kingdom dataset only by manga). Fourth row of Table II shows an improvement by 14.74% and 11.39% for the eBDtheque and Kingdom datasets respectively. It confirms the importance of using the tail position when available. This increase is higher for the eBDtheque dataset because European and American comics generally use a bigger tail than manga (less possible confusions for speaker association). Failure examples are illustrated Fig. 4. The first image (Fig. 4a) contains a crowd emitting two balloons, the crowd not being annotated as a comic character in the ground truth, no association could be computed. In Fig. 4b, a non-speaking comic character being in between the balloon and its speaker, the closest character does not correspond to the appropriate speaker for the biggest balloon of this panel. In Fig. 4c, the association failed because the speaker is not in the same panel as the balloon. In Fig. 4d, only one of two associations has been retrieved because the proposed method considers only one association per balloon.

### C. Results comparison

The overall performance increases when using a more precise anchor point selection for both evaluation scenarios (difference between *Method A* and *Method B*). The performance of previous element extractions has an important impact on the proposed method because more than 70% of the missed associations are due to missed balloons or missed comic characters. The impact of the tail information only becomes visible when prerequisites are properly extracted. As already mentioned, comic character extraction is not trivial, methods from the literature are not yet accurate enough to locate comic characters with a sufficient precision for our purpose.

## V. Discussion

The proposed method allows one-to-one speech balloon and comic character association. In case of speech balloons with multi-speakers, the proposed approach can be applied for each tail assuming that the tail extractor has the ability to propose several tails. Panel positions are good clues to compute the associations, in case they are hardly extractable by previous




(a) Unidentifiable speaker (b) In-between speaker

(c) Out-of-panel speaker (d) Multi-speaker

Fig. 4: Most frequent incorrect or missed associations (blue straight-lines). In this illustration, prerequisites are loaded from ground truth and their positions are represented by colored polygons. In red (panels), cyan (balloons) and yellow (comic characters). Image credits: acknowledgment section.

extractors (e.g. implicit panels), our method can still be used by considering the full image as being a single panel. In this approach, we measured the benefit of tail tip position but we believe that the tail direction is a complementary indication to consider. Tail direction can be integrated in the geometric graph has a property of each balloon vertex and matched with edge angles in order to guide the associations.

## VI. Conclusion

This paper presents a speech balloon and comic character association approach toward comics and manga understanding. The proposed method is composed by anchor point selection and geometric graph analysis. We also analyzed the impact of the prerequisites such as panel, speech balloon and comic character positions. The proposed approach is robust against missing panels and adaptive to different level of description of the balloons (e.g. presence of tail tip) and characters (e.g. body, face or mouth). In the future we plan to also consider tail direction in order to handle out-of-panel speakers and multi-speaker issues.

## Acknowledgment

This work was supported by the University of La Rochelle (France), the town of La Rochelle and the bilateral program PHC-SAKURA between Campus France and the Japan Society for the Promotion of Science (JSPS). We are grateful to all authors and publishers of comics and manga from eBDtheque and Kingdom datasets for allowing us to use their works.

## References

[1] Y. Mosallam, A. Abi-Haidar, and J.-G. Ganascia, "Unsupervised named entity recognition and disambiguation: An application to old french journals," in *Advances in Data Mining. Applications and Theoretical Aspects*, ser. Lecture Notes in Computer Science, P. Perner, Ed. Springer International Publishing, 2014, vol. 8557, pp. 12–23.

[2] L. Li, Y. Wang, Z. Tang, and L. Gao, "Automatic comic page segmentation based on polygon detection," *Multimedia Tools and Applications*, vol. 11042, pp. 1–27, 2012.

[3] 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*, ser. HCI'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 872–881.

[4] T. Tanaka, K. Shoji, F. Toyama, and J. Miyamichi, "Layout analysis of tree-structured scene frames in comic images." in *IJCAI'07*, 2007, pp. 2885–2890.

[5] 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*, vol. 5, no. 2, pp. 216–223, 2011.

[6] C. Rigaud, N. Tsopze, J.-C. Burie, and J.-M. Ogier, "Robust frame and text extraction from comic books," in *Graphics Recognition. New Trends and Challenges*, ser. Lecture Notes in Computer Science, Y.-B. Kwon and J.-M. Ogier, Eds. Springer Berlin Heidelberg, 2013, vol. 7423, pp. 129–138.

[7] 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.

[8] X. Pang, Y. Cao, R. W. Lau, and A. B. Chan, "A robust panel extraction method for manga," in *Proceedings of the ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 1125–1128.

[9] C. Ponsard, R. Ramdoyal, and D. Dziamski, "An ocr-enabled digital comic books viewer," in *Computers Helping People with Special Needs*. Springer, 2012, pp. 471–478.

[10] L. Li, Y. Wang, Z. Tang, and L. Gao, "Automatic comic page segmentation based on polygon detection," *Multimedia Tools Appl.*, vol. 69, no. 1, pp. 171–197, 2014.

[11] 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.

[12] 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.

[13] 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.

[14] 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, Aug. 2013, pp. 1240–1244.

[15] C. Guérin, C. Rigaud, K. Bertet, J.-C. Burie, A. Revel, and J.-M. Ogier, "Réduction de l'espace de recherche pour les personnages de bandes dessinées," in *Reconnaissance de Formes et Intelligence Artificielle (RFIA) 2014*, France, Jun. 2014.

[16] W. Sun and K. Kise, "Similar partial copy recognition for line drawings using concentric multi-region histograms of oriented gradients," in *Proceedings of the IAPR Conference on Machine Vision Applications*, ser. MVA2011, Nara, JAPAN, June 13-15, 2011.

[17] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[18] M. Iwata, A. Ito, and K. Kise, "A study to achieve manga character retrieval method for manga images," in *Proceedings of the 11th IAPR International Workshop on Document Analysis Systems (DAS2014)*, Apr. 2014, pp. 309–313.

[19] J. Pach, "Geometric graph theory," Cambridge University Press, Tech. Rep., 1999.

[20] 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, Aug. 2013, pp. 1145–1149.