An active contour model for speech balloon detection in comics

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

Dimosthenis Karatzas, Joost Van de Weijer
Computer Vision Center
Universitat Autònoma de Barcelona
E-08193 Bellaterra (Barcelona), Spain
{dimos, joost}@cvc.uab.es

***Abstract*—Comic books constitute an important cultural heritage asset in many countries. Digitization combined with subsequent comic book understanding would enable a variety of new applications, including content-based retrieval and content retargeting. Document understanding in this domain is challenging as comics are semi-structured documents, combining semantically important graphical and textual parts. Few studies have been done in this direction. In this work we detail a novel approach for closed and non-closed speech balloon localization in scanned comic book pages, an essential step towards a fully automatic comic book understanding. The approach is compared with existing methods for closed balloon localization found in the literature and results are presented.**

***Keywords*—active contour, multi-scale, non-closed contour, speech balloon, comics**

I. Introduction

Comic books are a widespread cultural expression and are commonly accepted as the "ninth art". They emerged in the United States in the early 20th century alongside with related media such as film and animation. Comics are a hybrid medium, combining textual and visual information in order to convey their narrative. The European market has traditionally been the biggest consumer of comic books, with the number of new paper comic titles in the EU quadrupling over the last decade experienced. Digitization combined with subsequent document understanding of comic books is therefore of interest, both in order to add value to existing paper-based comic heritage, but also to bridge the gap between the paper and electronic comic media. In comics content understanding, speech balloons present a lot of interest since they offer the links between the textual content and the comic characters providing information about the localization of the characters and the tone of speech (shape). Apart from being crucial for document understanding, balloon detection is also beneficial in applications such as comic character detection [1], content re-targeting [2], translation assistance and reading order inference [3].

The inside area of the balloons is usually light to improve readability, while they are surrounded by a black outline (contour). This contour has a particular shape that conveys the intonation of the text. In this study, we consider four usual shapes: rectangle, oval, cloud and peak which are the most frequent kinds of balloons. The contour is not always completely drawn, it may be "implied" due to contrast difference or other surrounding elements (see fig. 1c). In most of the cases, including when contours are implicit, the location of text is generally a good clue to guess where the balloon is. The problem of speech balloon outline detection can therefore be posed as the fitting of a closed contour around text areas, with the distinctiveness that the outline might not be explicitly defined in the image. For the examples given in figure 1, this would be a smooth contour with relatively constant curvature (fig. 1a), an irregular one with high local curvature (fig. 1b), and an implicit one with missing parts (fig. 1c).

a) Round b) Peaked c) Suggested

Fig. 1. Example of speech balloons. Image credits [4].

As far as we know, in the literature, only closed balloon extraction has been studied for comics. Those methods are based on either white blob detection (Arai [5]) or connected component extraction (Ho [6]) combined with heuristic filtering. Observing the heterogeneity of balloons, and considering the difficulty to manage "open" balloons, it appears necessary to use a dynamic and adaptive outline detection algorithm. Active contours appear to be suitable to the problem. The active contour framework was developed for delineating an object outline in an image. The algorithm attempts to minimize the energy associated to the current contour, defined as the sum of an internal and an external energy term. In this paper we propose two adaptations of the active contour theory to the domain of comic balloon detection. Specifically, we handle the case of balloons with missing parts or implicit contours, while we adopt a two-step approach to fit irregular outline types such as peak and cloud type balloons. To achieve this, we propose new energy terms making use of domain knowledge.

After reviewing the main ideas of active contour, section III and IV introduce the new aspects of this paper. We illustrate our method by showing the results of speech balloon detection in section V.

II. Active contours

The active contour [7] model is a deformable model, also known as snake, corresponding to a curve $\mathbf{v}(s) = [x(s), y(s)], s \in [0, 1]$, that moves through the spatial domain of an image to minimize the energy functional (eq. 1).

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

where α and β are weighting parameters that respectively control the snake's tension and rigidity, and $\mathbf{v}'$ and $\mathbf{v}''$ denote the first and second derivatives of $\mathbf{v}(s)$ with respect to s. This functional energy is also called E_{int} for internal energy. The external energy function E_{ext} is computed from the image so that it takes on its smaller values at the features of interest, such as boundaries [8]. One of the proposed energy functions by Kass [7] is eq. 2 which attracts the contour to edges with large image gradients.

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

Xu [8] proposed the Gradient Vector Flow external force to attract snake from further and handle broken object edges and subjective contours. Another extension was proposed by Cohen [9] to make the curve behave like a balloon which is inflated by an additional force. Finally, active contour in a multi-resolution context have been studied to speed up the process on multi-resolution image and the multi-resolution model itself [10].

III. Active contour for speech balloons

In this section we adapt the active contour framework to the domain of comics to detect speech balloons given a prior detection of text regions, introducing new energy terms based on domain knowledge about the relationship between text and balloons. For the discussion below, we consider that text has been already detected in the image, and we use the method of [11] for text detection.

The introduction of statistical shape knowledge has already been studied in the literature [12] but can not be applied here because of the lack of knowledge about the contour shape to detect. Hence we introduce a new energy term, denoted E_{text} (see eq.3) that conveys information about the relative placement of the balloon outline and the enclosed text.

$$E = E_{int} + E_{ext} + E_{text} \quad (3)$$

A. External energy

We consider edges as features of interest because we expect the speech balloon to be delimited by strong edges, at least partially. Edge detection is performed based on the Sobel operator (see fig. 2). The original Kass [7] external energy (see eq. 2) is appropriate for natural scene images with smooth gradients but not for comics that comprise uniform coloured regions and strong edges. In our case, we require that edges attract the snake from relatively far away (distances where the original edge gradient has already dropped to zero). The method of Xu [8] would be appropriate here, although we have decided to use the equivalent distance transform of the edge image instead for computational efficiency reasons. We therefore define the external energy function as:

$$E_{ext} = \gamma \min A(i, j) = \gamma \min \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (4)$$

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

Since it is not desirable for edges corresponding to text to attract the snake, any edges that fall within the detected text regions (see [11]) are removed before the distance transform is calculated and do not contribute to the external energy.

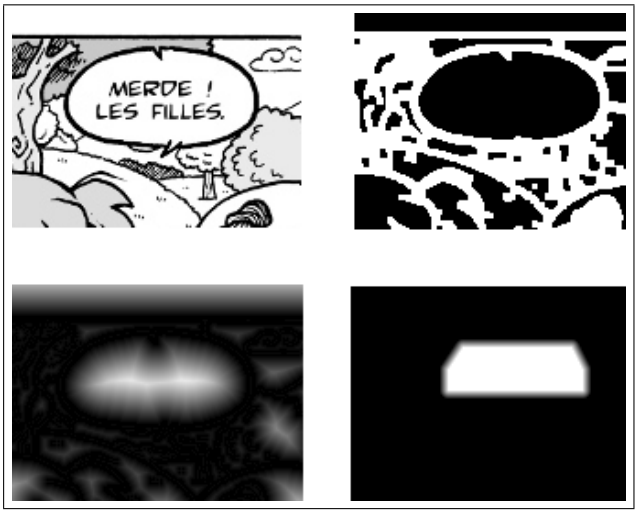

Fig. 2. Example of original image (top left) and its corresponding non-text edge detection (top right), E_{ext} energy (bottom-left) and E_{text} energy (bottom-right). In the bottom part, white corresponds to high energy.

B. Internal energy

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

E_{cont} forces the contour to be continuous by keeping points at equal distance, spreading them equally along the snake according to the average inter-point distance of the contour. It becomes small when the distance between consecutive points is close to the average, see eq. 5.

$$E_{cont} = \alpha|\bar{d} - \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}| \quad (5)$$

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

E_{curv} enforces smoothness and avoids oscillations of the snake by penalizing high contour curvatures (minimizing the second derivative). It becomes small when the angle between three points is close to zero, see eq. 6.

$$E_{curv} = \beta\left((x_{i-1} - 2x_i + x_{i+1})^2 + (y_{i-1} - 2y_i + y_{i+1})^2\right) \quad (6)$$

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

C. Text energy

The text energy E_{text} conveys domain specific knowledge about the relative locations of text areas and their associated

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

$$E_{text} = \begin{cases} \kappa \frac{N}{min_{j \in T} A(i,j)} & \text{if } A(i,j) > 0 \\ \kappa N & \text{else} \end{cases} \quad (7)$$

where j is a pixel in the text area T, N is an experimentally defined constant expressing the expected distance in pixel between the text area and the corresponding balloon boundary and κ is a weighting parameter that controls the contribution of E_{text} with respect to the other energy terms in eq. 3. Note when i is on the border of T, the distance $A(i,j)$ is equal zero and the energy becomes maximal as if it was inside T.

IV. Proposed method

In this section we detail how to localize speech balloons using active contours based on the definitions given above. First we generate the static external energy map E_{ext} for the whole image and then for each text area we compute the E_{text} energy. The internal energy E_{int} is calculated for each point of the snake before each iteration. We iteratively examine each point of the snake in a clockwise fashion and move it within a neighbourhood region of size M in order to minimize equation 3. This operation is repeated until no point moves in one turn (see algorithm 1). We perform a first smooth-contour approximation of the balloon boundary (low resolution) and then we fit the contour better to the balloon shape (high resolution) using the same algorithm.

Algorithm 1 Balloon detection

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

A. Active Contour initialization

In this study, we propose to rely on text localization to find speech balloons. Comics have several text categories depending on the purpose (e.g. speech, sound effect, illustration, narration), as we aim at detecting speech balloons, we base our work on the speech text localization. At this stage, any speech text detector can be used but our result highly relies on its performance. We use the algorithm presented in Rigaud et al [11] that reaches 75.8% recall and 76.2% precision for text line localization (mainly speech text). As we are interested in initializing the snake on a single text area for each balloon detection, we post-process the results of the text line detection [11] to group text lines into text area (paragraph) ones, according to two rules. First, we require that the candidate text lines to group have similar heights and second that the inter-line distance is smaller than their average text line height of the current paragraph (see fig. 3a). Given a resulting text area, we initialize the snake on its convex hull (see fig. 3b). Note that the convex hull of the text area also corresponds to the E_{text} maximal value border.

The initial number of points impacts the way that the snake moves and the precision of the final detection. During the first low resolution localization step, we perform a spaced equipartition of the points (see fig. 3c) to quickly localize the global shape avoiding unnecessary stops on image details. In the subsequent high resolution fitting stage, we add more intermediate points to fit the exact shape more precisely.

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

Fig. 3. Active contour initialization based on text area convex hull.

B. Low resolution contour localization

Following a two stage process, we first aim to obtain a rough localization of the balloons by fitting a smooth contour using a few contour points during initialization. The idea is to progressively push the snake away from the text area and towards the balloon boundary giving an increased weight to both the E_{ext} and E_{text} energy terms. If the balloon has an explicit boundary then E_{ext} will attract the snake to it. If there is no explicit contour close enough to attract the snake then the E_{text} term will push the snake to the suggested position of the balloon contour. Also the internal energies are important at this stage to maintain a certain rigidity of the snake. At the end of this step, we obtain a preliminary localization of the speech balloons, see fig. 4.

C. High resolution contour shape fitting

Figure 4 shows that the global shape of the top balloon has been detected although it is still far from a perfect fit because, as we can see on bottom part of the balloon, the snake was not able to describe the coarse parts of the boundary (e.g. peaks, tail). To achieve a better fitting, we increase the resolution of

Fig. 4. Example of low resolution contour detection (red line) for closed (left) and open (right) balloons.

the snake by adding new points between the current ones and by changing the weighting parameters of the energy function, and go through a second fitting process. At this stage, we relax the E_{curv} energy to make the snake fit to coarser parts of the boundary and we set E_{cont} strong enough to keep a regular inter-point distance all over the contour. Also, we reduce the E_{text} energy weight because at this step, the snake is already far from text and this term is not informative any more. This new configuration allows the snake to detect more coarse elements as shown in figure 5.

Fig. 5. Example of high resolution contour detection (red line) for closed (left) and open (right) balloons.

V. Experiments

In this section we evaluate the proposed method on a dataset of scanned comic books, and compare our results to other approaches, including a naive baseline.

A. Experimental setup

During the experiments the energy function was computed for each pixel within a neighbourhood of each point of the snake and the minimum score pixel becomes the new position of the snake point. We fixed the neighbourhood region M to 5x5 pixels and the E_{text} energy term's optimal distance parameter N at one half of the average text line height of the page, based on experimental validation. All the energies (eq. 4, 5 and 6) where normalised between zero and one by dividing by the maximum value of each energy terms. The snake was initialized with a variable number of points P so that the inter-point distance is less or equals to one quarter of the average text line height of the page. We based N and P on text height to be invariant to image definition.

The weighting parameters γ of the external energy term (E_{ext}), α and β of the internal energy terms (E_{int}) and κ for the text energy term (E_{text}) where defined based on validation as $\gamma = 0.2$, $\alpha = 0.1$, $\beta = 0.2$, $\kappa = 0.5$ for the localization step and $\gamma = 0.2$, $\alpha = 0.3$, $\beta = 0.4$, $\kappa = 0.1$ for the contour fitting step.

B. Dataset

We made experiments with the dataset eBDtheque [13] from which we selected 50 pages on 100 containing 453 speech balloons and 1547 text lines. In this dataset, balloon localization are given as orthogonal horizontal bounding boxes circumscribing either the balloon boundary when explicit or the contained text implicit. The tail of this balloon is ignored in the ground truth.

C. Performance evaluation

We evaluated our different contributions separately. First we measured balloon localization performance at bounding box level to highlight the benefits of both active contour theory and domain specific knowledge. Second, we performed pixel level evaluation on a smaller subset to show the ability of our method to fit balloon contour details.

1) Localization: The following results were obtained with the common evaluation measures of recall, precision and F_1 score at bounding box level. Recall (R) is the number of correctly detected balloons divided by the number of balloons in the ground truth. Precision (P) is the number of correctly detected balloons divided by the number of detected balloons. Each detected balloon D was compared to the corresponding ground truth one G with the nearest centroid, considering only one matching per ground truth balloon. We consider as correctly detected, balloons that overlap with a ground truth one more than 80% ($D \cap G > 0.8 * G$) and overflow the ground truth one less than 40% ($G - (D \cap G) < 0.4 * G$).

In order to provide a comparative analysis we attempted comparison to the methods of Arai [5] and Ho [6], which are based on connected component detection and filtering. Unfortunately, direct comparison to these methods is not possible as Arai's approach [5] is based on two rules specifically designed for Japanese manga comics with vertical text and Ho [6] is based on growing region segmentation which is not appropriate for open balloon detection neither. We also compared to the original active contour implementation proposed by Kass et al. [7] but because the initialization is not close enough to the edges, the internal energies make the snake retracts on itself.

Therefore, we compare our results to a baseline method (1) that considers as balloon any white connected components that overlaps at more than 10% with text regions. We also compare against the active contour with a distance transform based external energy described section III-A (2) and finally we add the E_{text} energy (3) from section. III-C. The results are presented in table I. For each method, we present two variants, one making use of ground truth localization for the text areas as seeds for balloon localization, and another making use the results of the automatic text localization method of [11].

TABLE I. Balloon localization recall and precision.

	Ground truth			Automatic		
Method	R (%)	P (%)	F_1	R (%)	P (%)	F_1
(1)	56.6	79.2	66.0	53.1	53.0	53.1
(2)	89.0	90.7	89.8	82.1	53.7	64.9
(3)	92.3	94.4	93.4	83.4	55.5	66.6

The baseline method (1) detects half of the balloons (about 56% recall) in this dataset as it is not able to detect open balloons or balloons containing little text. However, it has two advantages in comparison with the proposed active contour based method. First, as the a connected component is considered as balloon, when correct, the precision at the pixel level is maximal. Second, it is faster to compute. The results using active contour with distance transform (2) shows a significant improvement, thanks to the active contour theory that detect much more balloons whether open or closed than connected component based methods. Doing just the low resolution localization step, or continuing to include the high resolution fitting step does not cause any difference for our method (3) under this evaluation scheme, as the evaluation is performed at the level of bounding boxes overlapping.

2) Contour fitting: To evaluate the benefits of the second stage we propose, we repeated the evaluation using pixel level ground truth, on a small subset of three comic pages (24 balloons). The results are shown in table II. Note that for this experiment we selected three pages where ground truth and automatic text detection give the same results.

TABLE II. BALLOON SHAPE DETECTION RECALL AND PRECISION.

	Single-stage		Two-stage	
	R (%)	P (%)	R (%)	P (%)
Page 1	91.3	86.5	95.9	90.1
Page 2	79.0	95.9	78.4	94.8
Page 3	93.4	92.8	97.7	88.7

Table II shows higher score for the two-stage variant for the page 1 and 3. These two pages contain mainly closed balloons, we see here that the second stage improves the accuracy of the detection of closed balloons. In the case of implied balloon boundaries, as in page 2, the second stage is not resulting in any improvement as there is no extra local information that can assist in the fitting. In this case the results mainly depends of the low resolution detection. Note, processing time was about 10 seconds for a 300DPI A4 image on a regular machine.

VI. DISCUSSION

The presented method highly depends on the active contour initialization success. In this study, we relied on speech text detection as we assume it is the most common feature that balloons include, while past experiments have shown that its accurate detection is feasible and stable. A side-effect of this choice is that the text line detector we used was not able to detect balloons that contain other contents (e.g. drawings, punctuation). We believe there is room for improvement of the different energy terms we used. For example, one could use the Gradient Vector Flow proposed by Xu [8] for the external energy, especially in the case of missing data balloon boundaries. On the other hand, the ground truth of implicit balloons is at best questionable as the exact localization of the balloon is quite subjective. An way to circumvent this problem could be to either define the boundary in a flexible way, or directly define ground truth at the pixel level. All the materials for reproducing and comparing the results presented in this paper are publicly available through http://ebdtheque.univ-lr.fr/references.

VII. CONCLUSION

We have proposed and evaluated a new active contour based method to accurately localize open and closed speech balloon in comic book. The proposed approach relies on text detection and prior knowledge to fit balloon contour at different resolutions. The evaluation shows 92.3% recall and 94.4% precision using ground truth text and 83.4% recall and 55.5% precision using an automatic text detector. Further effort has to be made to define open balloons ground truth.

ACKNOWLEDGMENT

The authors would like to thank Clement Guérin for his work about the evaluation framework we used to evaluate this work section V-C. This work was supported by the European Doctorate founds of the University of La Rochelle, European Regional Development Fund, the region Poitou-Charentes (France), the General Council of Charente Maritime (France), the town of La Rochelle (France) and the Spanish research projects TIN2011-24631, RYC-2009-05031.

REFERENCES

[1] W. Sun and K. Kise, "Similar Manga Retrieval Using Visual Vocabulary Based on Regions of Interest," *2011 International Conference on Document Analysis and Recognition*, pp. 1075–1079, 2011.

[2] Y. Matsui, T. Yamasaki, and K. Aizawa, "Interactive Manga retargeting," *ACM SIGGRAPH 2011 Posters on - SIGGRAPH '11*, p. 1, 2011.

[3] C. Guérin, "Ontologies and spatial relations applied to comic books reading." in *PhD Symposium of Knowledge Engineering and Knowledge Management (EKAW)*, Galway, Ireland, 2012.

[4] Cyb, *Bubblegôm Gôm vol. 1, pp. 3*. Goven, France: Studio Cyborga, 2009, vol. 1, p. 3. [Online]. Available: http://bubblegom.webcomics.fr/page/les-filles-de-troisieme

[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] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[8] C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow." *IEEE Transactions on Image Processing*, vol. 7, no. 3, pp. 359–369, 1998.

[9] L. D. Cohen, "On active contour models and balloons," *CVGIP: Image Understanding*, vol. 53, no. 2, pp. 211–218, Mar. 1991.

[10] B. Leroy, I. Herlin, and L. Cohen, "Multi-resolution algorithms for active contour models," in *ICAOS'96*, ser. Lecture Notes in Control and Information Sciences, M.-O. Berger and D. et al., Eds. Springer Berlin Heidelberg, 1996, vol. 219, pp. 58–65.

[11] 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.

[12] D. Cremers, F. Tischhäuser, J. Weickert, and C. Schnörr, "Diffusion snakes: introducing statistical shape knowledge into the mumford-shah functional," *International Journal of Computer Vision*, vol. 50, pp. 295–313, 2002.

[13] 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.