Automatic Text Localisation in Scanned Comic Books

Christophe Rigaud[1], Dimosthenis Karatzas[2], Joost Van de Weijer[2], Jean-Christophe Burie[1], Jean-Marc Ogier[1]

[1]*Laboratory L3i, University of La Rochelle, Avenue Michel Crépeau 17042 La Rochelle, France*

[2]*Computer Vision Center, Universitat Autnoma de Barcelona, E-08193 Bellaterra (Barcelona), Spain*

{*christophe.rigaud, jean-marc.ogier , jean-christophe.burie*}*@univ-lr.fr,* {*dimos, joost*}*@cvc.uab.es*

Keywords: Text localization, comics, text/graphic separation, complex background, unstructured document

Abstract: Comic books constitute an important cultural heritage asset in many countries. Digitization combined with subsequent document understanding enable direct content-based search as opposed to metadata only search (e.g. album title or author name). Few studies have been done in this direction. In this work we detail a novel approach for the automatic text localization in scanned comics book pages, an essential step towards a fully automatic comic book understanding. We focus on speech text as it is semantically important and represents the majority of the text present in comics. The approach is compared with existing methods of text localization found in the literature and results are presented.

1 INTRODUCTION

Comic books constitute an important cultural heritage asset in many countries. In this work we detail a novel method for the automatic text localization in scanned comic books written in Latin script, an essential step towards a fully automatic comic books understanding.

Most of the text conveys communication between characters and is written into speech balloons. Other categories of text concern the narrative text and the onomatopoeias. Narrative text describes a scene, provides background to the story, or explains what is not visible in the frames. The onomatopoeias represent sounds in a textual way or a sequence of symbols to express a mood. The exact usage of text is a matter of style and depends on the author. Nevertheless, speech and narrative text is generally written in a salient way, most of the time in black ink over clear background. Furthermore, comics text can be handwritten and/or typewritten depending on the style of the comic. In both cases, calligraphy and typography are letter-spaced for readability purposes.

Both handwritten and typewritten text recognition are open areas of research. The focus of this work is not to recognize the text but to localize it and separate it from the graphical content. In comics, as in many other unstructured documents, text recognition depends on text localization. The text may be everywhere within the drawing area which produces a lot of noise and false detections if we submit it directly to an OCR process. Moreover, text in comics presents a lot of variations (e.g. stroke, orientation, colour, width, height) that can cause many issues at both the localization and recognition processes. Text localization provides text-only images to a recognition system to improve recognition accuracy and reduce computing time. Moreover, text localization is also important because it is related to many parts of the comics content (e.g. speech balloon, characters and panels) and defines the reading order (Tsopze et al., 2012). Detecting these parts and establishing their relationship is crucial towards complete comics understanding. In addition, text localization in comics opens up several interesting applications such as image compression (Su et al., 2011), OCR training and also translation, speech synthesis and retargeting to different mobile reading devices (Matsui et al., 2011).

Text localization in real scenes and video frames is an active research topic (Jung et al., 2004). However, applying existing text detection methods to comics would fail because the nature of comic documents is different. Comics being unstructured graphical documents, combine the difficulties of both domains, making the task of text localization especially challenging. On one hand, they differ from classical documents in that they comprise complex backgrounds of a graphical nature. Furthermore, they belong to the class of non-structured documents meaning there is no regular structure present for the prediction of text loca-

tions and no layout method applicable. Text localization in complex images has been previously studied in scene images (Weinman et al., 2009; Epshtein et al., 2010; Neumann and Matas, 2012) and (Wang and Belongie, 2010; Meng and Song, 2012), video sequences (Kim and Kim, 2009; Zhong et al., 2000; Shivakumara et al., 2009) and digital-born images (Web and email) (Hu and Bagga, 2004; Mori and Malik, 2003; Karatzas and Antonacopoulos, 2007). Text localization in unstructured documents has been studied in teaching boards (Oliveira and Lins, 2010) for example. However, text localization in documents which are both unstructured and have complex background has received relatively little attention (Clavelli and Karatzas, 2009).

To solve the particular problems which are provoked by the combination of complex background and unstructured documents, we propose a new text localization method. We improve the initial segmentation step to cope with complex backgrounds. Furthermore, we adapt the line extraction method to cope with unstructured documents. To evaluate our method we propose a new benchmark dataset for text localization in comic documents which will be made publicly available in few months. In our results we show that our method outperforms existing text localization methods applied on comics. Moreover, we believe that the proposed method generalizes to other unstructured images with complex backgrounds such as maps, posters and signboard images.

This paper is organised as follows. Section 2 presents an overview of text detection from comics. Sections 3 and 4 present the proposed method. Finally, sections 5 and 6 shows some experiments and conclude this paper.

2 RELATED WORK

Text/graphic separation was studied for different applications as the analysis of newspaper images, administrative documents, cheques and graphical documents such as maps and engineering drawings. There is little work published on comics text/graphic separation and text localization. Bottom-up approaches use connected component (CC) algorithms which depend on a segmentation step. (Su et al., 2011) use Sliding Concentric Windows as text/graphic separation then use mathematical morphology and an SVM classifier to classify text from non-text CC. As the morphological operations are performed with a fixed mask size, the method is orientation and resolution dependent. (Rigaud et al., 2012) make use of "the median value of the border page pixels" to binarize the image, extract CC and then classify them into "noise", "text" or "frame" (based on CC heights). This method assumes that the page always contains text and that the text background colour is similar to the paper background. A top-down approach starting with speech balloon (white blob) detection following by mathematical morphology operations to find lines of text is proposed by (Arai and Tolle, 2011). This method relies on balloon detection which is limited to closed balloons in this case. Another top-down approach that defines a sliding window of a character size to detect letters is suggested by (Yamada et al., 2004).

The contributions of this paper come at different levels. First, an adaptive comics page segmentation method is presented. Second, a text/graphic separation algorithm is proposed based on the local contrast ratio. Finally, a grouping method is used to create line hypotheses from detected possible text components as a final verification step.

3 TEXT LOCALIZATION IN COMPLEX BACKGROUND DOCUMENTS

In this section we propose several adaptations to the standard text localization pipeline to overcome the problems introduced by complex background.

3.1 Adaptive segmentation

Segmentation is a crucial step in many text localization methods. Comic speech text is made of strokes, generally black on white, that we need to segment. A perfect text segmentation would result to individual characters represented by single connected components. The complex background of comic documents complicates this step. Typical thresholding methods would fail to segment the text because text is drawn with the same style of strokes as many other graphical elements in the page.

Therefore, we propose an adaptive segmentation method. For a single page we assume that the text background brightness is similar around all the characters of the same page. However, in our case, the optimal segmentation threshold differs for every single page of comics depending on the background colour of the text areas. The method is based on the observation that choosing the threshold too low, as well as choosing the threshold too high leads to an excess of connected components (CC), as it can be observed in the figure 1.

This phenomenon is intrinsic to the nature of

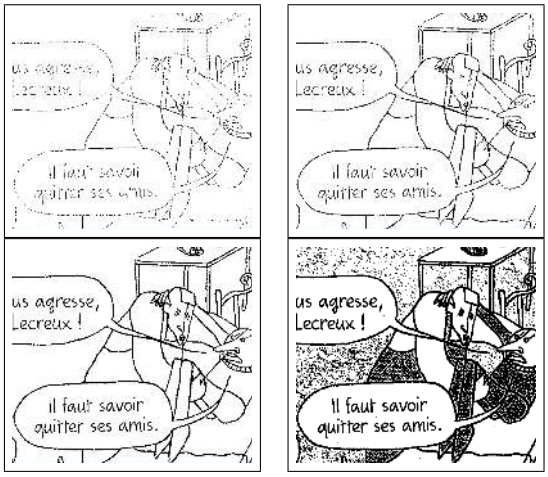

Figure 1: Segmentation at different threshold levels from the lower top-left to the higher bottom-right (threshold = 50, 100, 150, 200). We observe that the number of CC increases when the dark lines are cut and also when background start to appear as salt and paper noise. Source: (Roudier, 2011).

comics (and other graphical documents), as due to the design process they contain textured areas and loosely connected strokes that give rise to merged components at different thresholds. This is intensified by the digitisation and image compression process that adds further noise to the page.

Our method, minimum connected components thresholding (MCCT), automatically finds the right threshold by computing the number of CC for different threshold levels in a range of values from th_{min} to th_{max} and selecting the one that produces the minimum number of CC (in a 8 bits graylevels image). Then we find the first minimal number of CC. See example on figure 2. Note, that the optimal threshold is correctly predicted by MCCT as it corresponds to the bottom-left image in figure 1.

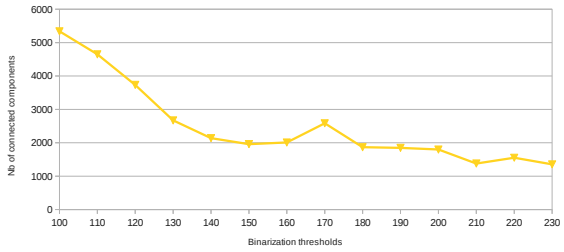

Figure 2: Example of number of CC function of the segmentation threshold. In this case, $th_{min} = 100$, $th_{max} = 230$ and our algorithm found 150 as optimal threshold threshold.

As an alternative segmentation method, we also tried to extract text CC using the Maximally Stable Extremal Region (MSER) algorithm (Matas et al., 2002) nevertheless, this algorithm produces an excess of false detections as in general a lot of the graphical elements are equally stable as the text. We also try the well known Otsu (Otsu, 1979) algorithm (see section 5).

3.2 Text/graphics separation

After the segmentation step, the CC may correspond to graphics, single letters or a sequence of letters if some letters are connected together (see figure 3). The objective of this section is to separate the last two categories from the first one. Note that the merged letters will not affect our results (if they are part of the same word) as we are aiming to text line localization.

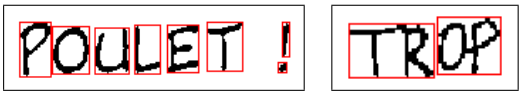

Figure 3: Example of connection between letters. On the left a word with only detached letters, on the right a word detected as two pairs of attached letters because of the handwriting.

We propose a number of rules, applied sequentially, to separate graphics from text. Due to the wide variety of text usage in comic albums, the method should be size, translation, rotation, scale and contrast invariant. In addition, it should be robust to text-less pages which may appear randomly within an album. From all the components extracted by the CC algorithm, we use three filtering rules to select only the ones corresponding to textual elements.

- Rule 1: we compare the standard deviation of the graylevels of each CC bounding boxes (see figure 4) with the contrast ratio of the page to make a hight/low local contrast CC separation. We assume a Gaussian distribution but other distribution are under study. The contrast ratio of the page is the absolute difference between the minimum and the maximum graylevels found on the page. In order to avoid artefacts pixel values, we apply a median filtering as preprocessing.

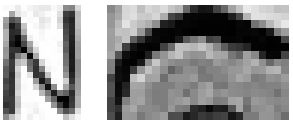

Figure 4: Example of mean (μ) and standard deviation (σ) values of two CC bounding boxes. For the letter N, $\mu, \sigma = [167, 84]$, for the eyebrow $\mu, \sigma = [110, 57]$.

- Rule 2: we apply a topological filtering that consists in removing all the CC that overlap with others assuming that a letter CC can't contain another CC (similar to (Thotreingam Kasar et al., 2007)). In our study we consider only the black on white CC. (see figure 5).

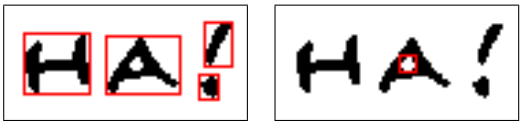

Figure 5: On the left an example of black on white CC bounding box (red rectangle) and the right a white on black.

- Rule 3: we check that each letter l_i is surrounded by other letters l_j (with $_j \in [0,n]$) which is intrinsic to the design of text (grouping characters in words and text lines). To do so, we check for similar in height components in an area equal to the size of the component's bounding box extended to its left/right/top/bottom directions (figure 6). For instance, l_i is similar in height to l_i if $abs(l_i.height - l_j.height) < a * l_i.height$ (a =similarity offset).

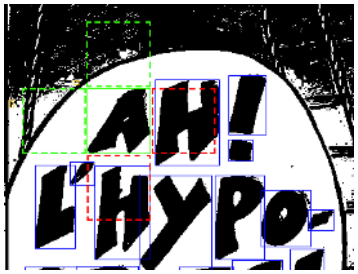

Figure 6: Example of the letter validation method (rule 3). The area of the bounding box of the letter "A" is extended in four directions around the letter "A" to check if any other letter (blue rectangles) overlaps with it. In this case, there are two overlapping components (red dashed rectangles) at the right and the bottom of the letter "A". The component is thus accepted as a possible letter.

- Rule 4: we check all the CC bounding boxes pairs that overlap more than $b\%$ each other. We remove the biggest from each pairs.

The four rules exploit contrast, topology and text structure to classify the CCs as being either graphics or text.

4 TEXT LOCALIZATION IN UNSTRUCTURED DOCUMENTS

In this paper we consider the term *unstructured documents* to define documents in which text can be arbitrarily placed and oriented within the page. In this section we propose a method for text localization in such documents.

The gap between letters (or attached letters) and lines can vary significantly with the rotation. In fact, handwriting generates many artefacts such as lack of good alignment, mergers between letters and mergers between text lines. We propose a method that handles the two first aforementioned artefacts considering only the height of the CC.

We first look for the first letter of each text line. A letter is considered first if it is positioned on the left, on the same horizontal line and if there is no intersection found with any other letters at a distance equal to the letter height. Then the other letters on the right are added by checking their relative horizontal and vertical positions. For this purpose, we defined two conditions that are auto-adaptive to each letter (see figure 7):

- The horizontal inter-letter distance d should be smaller than the maximal height of the letters ($d < Max(h1,h2)$);
- The vertical alignment is considered as correct if the horizontal position of the centre of the next letter $c2$ passes trough the first letter ($y_{min}(letter_1) > c2.x$ and $y_{max}(letter_1) > c2.x$);

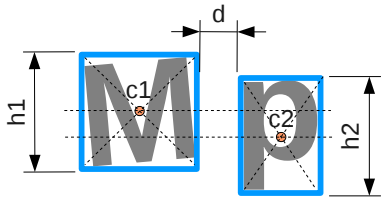

Figure 7: Letter horizontal and vertical positions variables (on the left $letter_1$, on the right $letter_2$). The bounding boxes are drawn in blue and the geometric centres ($c1, c2$) in orange.

This is similar in principle to (Clavelli and Karatzas, 2009) but adapted to take into account the horizontal alignment of text lines. Our method does not use the letter width as we never know how many letters correspond to the CC (due to possible letter mergers).

5 EXPERIMENTS AND RESULTS

In this paper we have proposed a method for text localization in unstructured documents with complex background. In these results we compare our results to other text localization methods, which were either designed for unstructured documents (Tombre et al., 2002), complex backgrounds (Neumann and Matas, 2012).

For the experiments, we applied a 3x3 median filtering on the whole image in order to smooth pixels values and then the MCCT was performed with a threshold computed from a range $[th_{min}, th_{max}] = [100, 230]$. The range was defined from the experiments as there is no enough information at lower or higher segmentation levels. We ignored CC smaller than 6 pixels because typically 6 pixels are needed to make letters fully recognizable by subjects (Wright, 2002). The hight/low local contrast CC separation (rule 1) was computed from a threshold of 50% based on experiments. For the text grouping (rule 3), the height similarity offset were defined at $a = 1/2$ as we assume that uppercase letters are twice bigger than lowercase letter. Finally, the maximum overlapping percentage was fixed at 30% (rule 4) based on experiments. Results are shown figure 8. In some cases we couldn't reach 100% recall because the proposed method is not fitted for the graphical text (e.g. sound effects, onomatopoeias, page title) and bright over dark text which are in the ground truth with no distinction.

An alternative segmentation method to the graylevel segmentation as one-pass colour segmentation algorithm which creates 8-connected components based on colour similarity was used (see section 5.3).

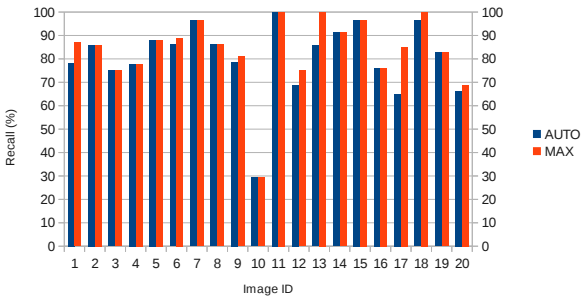

Figure 8: This figure shows the maximum recall obtained manually (MAX) and the adaptive segmentation (AUTO) for a sample of images of the dataset. Image 10 contains more than 60% of bright text over dark background which is not detected by our algorithm.

5.1 Dataset

The dataset used was composed of comics pages sourced from European publisher which contain 1700 text lines in total. The pages were selected from 15 albums belonging to different authors in order to be as representative as possible. They are in different format (e.g. A4, A5, single strip) and resolution (e.g 100 - 300 DPI). The dataset and the ground truth will be made available in the pubic domain[1].

5.2 Ground truth

In order to evaluate our algorithm, we created a ground truth information for all the dataset pages using a tool developed by the laboratory L3i[2]. As we consider text localization only, an easy way would be to draw a rectangle (bounding box) around each paragraph but the problem is that in comics, paragraph level bounding boxes may overlap with a large non-text part of the image (see figure 9). Hence we decided to produce ground truth at line level which is considerably better than paragraph level (although of limited accuracy in the case of text height variations and rotation). We intend to investigate word and letter levels in a future work.

Figure 9: Ground truth bounding boxes (translucent green rectangles overlapping text) at paragraph level (left part) which overlap graphic elements at the bottom-right part and blank areas at the extremities of lines. At line level (right part) which is a lot more accurate than paragraph level. Sources: (Cyb, 2009).

Note that the ground truth was created for all types of text with no distinction between sound effects (onomatopoeias), graphic text, speech text and narrative text (see figure 10). As we focus on speech text in this paper, we do not expect to reach 100% text detection for some images.

5.3 Evaluation

As far as we know, there is no similar evaluation scheme defined for comics in the literature, therefore we employ a well known real scene text localization method. We used (Wolf and Jolion, 2006) to evaluate our results following the approach of the ICDAR

[1] http://ebdtheque.univ-lr.fr

[2] ToonShop: http://l3i.univ-larochelle.fr/eBDtheque

Figure 10: Different type of text labelled in the ground truth. Sound effect (left), graphic text (center) and narrative text (right).

competition 2011 (Karatzas et al., 2011). We used the area precision thresholds proposed in the original publication: $t_p = 0.4$ and we decreased the area recall threshold from $t_r = 0.8$ to $t_r = 0.6$ in order to be more flexible with accent and punctuation miss detections (e.g. À, È, É, ..., !, ?) that enlarge the bounding boxes of line level ground truth. No splits and merges penalizations were considered, as text detection at either the word or text-line level is considered as correct. The table bellow shows a comparison with different segmentation and text/graphic separation methods from the literature, for more details on our results see figure 12.

Segment.	**Text/graphic sepa.**	**R (%)**	**P (%)**
(Neumann and Matas, 2012)		12.56	30.19
Colour	Proposed	15.69	6.92
Proposed	(Tombre et al., 2002)	74.18	61.25
Otsu	Proposed	75.14	64.14
Proposed	Proposed	**75.82**	**76.15**

Figure 11: Recall (R) and precision (P) results for different method combinations.

Because comics are really specific documents, real-scene text detection (Neumann and Matas, 2012) detect very few text lines. The best results we reached with a method from literature is a text/graphic separation method design for documents (Tombre et al., 2002) based on our adaptive segmentation (MCCT).

The combination of the proposed adaptive segmentation (MCCT) following by our text/graphic separation based on local contrast ratio beats the best method tested from the literature of 0.68% recall and 12.01% precision (see detail on figure 12). All the dataset was proceeded in less than 5 minutes on a regular machine with a 2.50GHz CPU and 8GB RAM.

We have made available our text detection method online[3] where images can be uploaded at any time and results are calculated and displayed in real time.

[3]http://www.christophe-rigaud.com/en/tools/text-detection-in-comics/

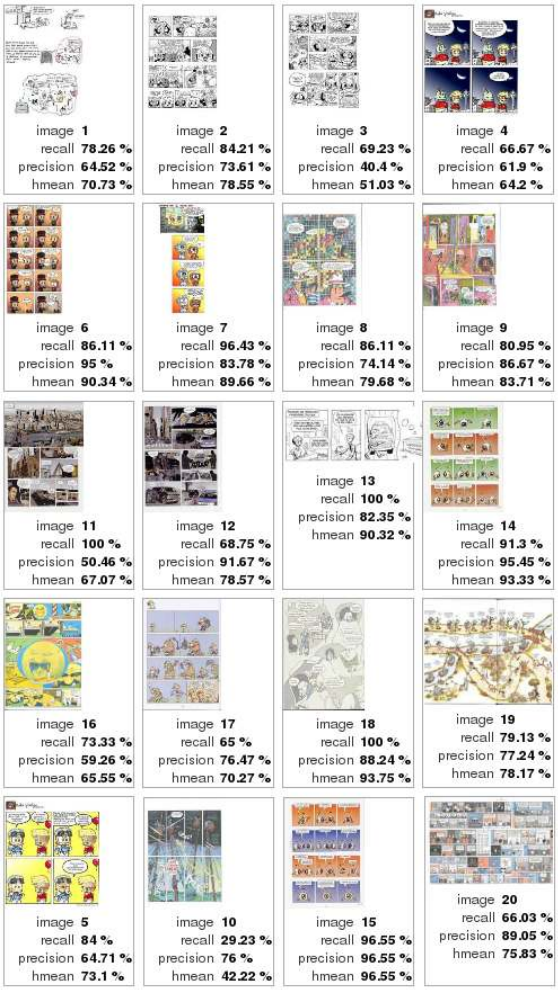

Figure 12: Result of our methods for 20 images of the dataset.

6 CONCLUSION AND PERSPECTIVES

We have proposed and evaluated a new method to localize text lines in comics books. The proposed approach is composed by a minimum connected components thresholding (MCCT) to segment text candidate connected components, a text/graphic separation based on contrast ratio and text line detection. The evaluation shows that more than 75.8% of the text lines are detected with 76% precision. However, the current method focuses on speech text, and further effort has to be made to detect other types of text such as graphic text. Text localization was a first step towards automatic comic book understanding, our future work will build on text detection results to look into speech balloon detection.

7 ACKNOWLEDGEMENTS

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

Arai, K. and Tolle, H. (2011). Method for real time text extraction of digital manga comic. *International Journal of Image Processing (IJIP)*, 4(6):669–676.

Clavelli, A. and Karatzas, D. (2009). Text segmentation in colour posters from the spanish civil war era. In *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, ICDAR '09, pages 181–185, Washington, DC, USA. IEEE Computer Society.

Cyb (2009). *Bubblegôm*. Studio Cyborga, Goven, France.

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

Hu, J. and Bagga, A. (2004). Categorizing images in web documents. *Multimedia, IEEE*, 11(1):22 – 30.

Jung, K., Kim, K. I., and Jain, A. K. (2004). Text information extraction in images and video: a survey. *Pattern Recognition*, 37(5):977 – 997.

Karatzas, D. and Antonacopoulos, A. (2007). Colour text segmentation in web images based on human perception. *Image and Vision Computing*, 25(5):564 – 577.

Karatzas, D., Mestre, S. R., Mas, J., Nourbakhsh, F., and Roy, P. P. (2011). Icdar 2011 robust reading competition - challenge 1: Reading text in born-digital images (web and email). *International Conference on Document Analysis and Recognition*, 0:1485–1490.

Kim, W. and Kim, C. (2009). A new approach for overlay text detection and extraction from complex video scene. *Image Processing, IEEE Transactions on*, 18(2):401 –411.

Matas, J., Chum, O., Urban, M., and Pajdla, T. (2002). Robust wide baseline stereo from. In *In British Machine Vision Conference*, pages 384–393.

Matsui, Y., Yamasaki, T., and Aizawa, K. (2011). Interactive manga retargeting. In *ACM SIGGRAPH 2011 Posters*, SIGGRAPH '11, pages 35:1–35:1, New York, NY, USA. ACM.

Meng, Q. and Song, Y. (2012). Text detection in natural scenes with salient region. In *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*, pages 384 –388.

Mori, G. and Malik, J. (2003). Recognizing objects in adversarial clutter: breaking a visual captcha. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–134 – I–141 vol.1.

Neumann, L. and Matas, J. (2012). Real-time scene text localization and recognition. *Computer Vision and Pattern Recognition*, pages 1485–1490.

Oliveira, D. M. and Lins, R. D. (2010). Generalizing tableau to any color of teaching boards. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, ICPR '10, pages 2411–2414, Washington, DC, USA. IEEE Computer Society.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66.

Rigaud, C., Tsopze, N., Burie, J.-C., and Ogier, J.-M. (2012). Robust frame and text extraction from comic books. *Lecture Note for Computer Science GREC2011*, 7423(19).

Roudier, N. (2011). *LES TERRES CREUSEES*, volume Acte sur BD. Actes Sud.

Shivakumara, P., Phan, T., and Tan, C. L. (2009). A robust wavelet transform based technique for video text detection. In *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on*, pages 1285 –1289.

Su, C.-Y., Chang, R.-I., and Liu, J.-C. (2011). Recognizing text elements for svg comic compression and its novel applications. In *Proceedings of the 2011 International Conference on Document Analysis and Recognition*, ICDAR '11, pages 1329–1333, Washington, DC, USA. IEEE Computer Society.

Thotreingam Kasar, Jayant Kumar, and Ramakrishnan, A. G. (2007). Font and Background Color Independent Text Binarization. In *Intl. workshop on Camera Based Document Analysis and Recognition (workshop of ICDAR)*, pages 3–9.

Tombre, K., Tabbone, S., Plissier, L., Lamiroy, B., and Dosch, P. (2002). Text/graphics separation revisited. In *in: Workshop on Document Analysis Systems (DAS*, pages 200–211. Springer-Verlag.

Tsopze, N., Guérin, C., Bertet, K., and Revel, A. (2012). Ontologies et relations spatiales dans la lecture d'une bande dessinée. In *Ingénierie des Connaissances*, pages 175–182, Paris.

Wang, K. and Belongie, S. (2010). Word spotting in the wild. In Daniilidis, K., Maragos, P., and Paragios, N., editors, *Computer Vision ECCV 2010*, volume 6311 of *Lecture Notes in Computer Science*, pages 591–604. Springer Berlin / Heidelberg.

Weinman, J., Learned-Miller, E., and Hanson, A. (2009). Scene text recognition using similarity and a lexicon with sparse belief propagation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(10):1733 –1746.

Wolf, C. and Jolion, J.-M. (2006). Object count/area graphs for the evaluation of object detection and segmentation algorithms. *Int. J. Doc. Anal. Recognit.*, 8(4):280–296.

Wright, S. L. (2002). Ibm 9.2-megapixel flat-panel display: Technology and infrastructure.

Yamada, M., Budiarto, R., Endo, M., and Miyazaki, S. (2004). Comic image decomposition for reading comics on cellular phones. *IEICE Transactions*, 87-D(6):1370–1376.

Zhong, Y., Zhang, H., and Jain, A. (2000). Automatic caption localization in compressed video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(4):385 –392.