Toward speech text recognition for comic books

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

Toward speech text recognition for comic books

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

ABSTRACT

Speech text in comic books is placed and written in a particular manner by the letterers which raises unusual challenges for text recognition. We first detail these challenges and present different approaches to solve them. We compare the performances of generic versus specifically trained OCR systems for typewritten and handwritten text lines from French comic books. This work is evaluated over a subset of public (eBDtheque) and private (Sequencity) datasets. We demonstrate that generic OCR systems perform best on typewritten-like and lowercase fonts while specifically trained OCR can be very powerful on skewed, uppercase and even cursive fonts.

CCS Concepts

•Information systems → Content analysis and feature selection;

Keywords

Handwritten text recognition; comics image analysis.

1. INTRODUCTION

Comic books are part of the cultural heritage of many countries and their massive digitization allows information retrieval. Text is one of the crucial information to retrieve in order to index comic book content. There are different type of text present in comics. We focus on the most frequent one which is speech text but there are other types of text such as title, caption, illustrative and drawing text (Figure 1). This paper highlights previous work about text recognition applied to comics and propose new perspectives towards typewritten and handwritten speech text recognition.

Text recognition in comics is really challenging because it includes most of the difficulties from text recognition in document analysis domain if we consider high variability of the text types that compose the comics. From typewritten to handwritten, free-form text in uniform to complex background including image noise, text deformation and overlapping. A desirable research contribution has been made in the field of text recognition considering scanned document pages from printed and handwritten text books. The main published studies on text recognition and identification on comics are discussed below.

Figure 1: Diversity of text types. Image credits: eBDtheque dataset.

Automatic text extraction and text recognition from speech balloon considering digital English comics was investigated [7]. In their investigation, a region based text extraction method was applied initially and furthermore, two sub-approaches: connected component and edge-based technique were introduced. Connected component labeling-based algorithm was applied to remove the noises from RGB images for detecting the connected components in the image. Connected component-based methods were applied by grouping small components into successively larger ones until all regions were identified in the image. In their research, color digital English comic image was taken as input and RGB band values were applied to input image for band selection. In recognition phase, the process of Optical Character Recognition (OCR) was divided into: segmentation, correlation and classification steps. In segmentation step, each text character was cropped and in the correlation step, cropped characters were matched with the datasets. In the classification process, text images was recognized considering the matching process of crop characters with the datasets. Finally, the text from OCR output was stored in text file for convenient use. Another method using OCR output have been proposed by Ponsard [6], it is part of an end-to-end process and focuses on speech text of a single French typewritten font for which an OCR system have been trained for.

Recently, a comic text recognition method was investigated [3]. In the work, Manga images were used for text extraction and text recognition process. A median filtering-based technique was considered in the pre-processing stage

MANPU '16 Cancun, Mexico

DOI: http://dx.doi.org/10.1145/3011549.3011557

for noise removal. A connected component labeling-based algorithm was taken into account for balloon detection in comic pages and subsequently the OCR was used for text recognition. The OCR process, described in their study was divided into: segmentation, correlation and classification steps in recognition phase. Text lines, words and characters were segmented before feeding characters into the OCR system. A desirable recognition rate at character level was achieved in the experimentation.

An identification technique considering text images in comic books was presented in our previous paper [10]. In this investigation, an attempt was made to explore a comic text identification technique of speech balloon to feed the identified text into the appropriate OCR system. Latin and Bengali comic text lines have been considered for identification. Two different local features, namely, Scale Invariant Feature Transform (SIFT) and Multi-scale Block Local Binary Pattern (MBLBP) were considered in Spatial Pyramid Matching (SPM) domain. The support vector machine (SVM)-based classification technique was considered in the experiment for text identification. A desired identification accuracy of 98.30% was achieved in the experiments.

In another paper [11], a method to recognize text elements in comic was proposed. The proposed method uses a sliding concentric windows (SCW) and support vector machine (SVM) based approach to identify text regions. Subsequently, OCR was applied to recognize text elements in speech balloons. Instead of encoding the text regions as vectors, the text elements were embedded in the SVG file along with coordinate values.

From the literature surveyed, it can be noted that an impressive progress of research is achieved in some areas of comics, such as speech balloon detection, text region localization, comic text extraction, and comic character recognition considering European, American and Manga comic books. Conversely, few investigations has been considered in text recognition in the field of comic image analysis.

In the next section, we compare typewritten and handwritten speech text in comics. Section 3 proposes two different processes, one for typewritten and one for handwritten text. Section 4 details the experiments we carried out on two datasets and Section 5 concludes this work.

2. TYPEWRITTEN VS HANDWRITTEN

Typewritten and handwritten texts are always appearing in intermixed way in several kinds of documents. Likewise, the comic books are also composed by these two different types of text. Usually, comic books consist of a wide variety of content types and the text part and comic characters are properly interlinked in a speech balloon. The textual part generally resides inside a speech balloon of comic books. Different types of text in terms of font size can be found inside balloon. It is observed that typewritten text characters have a relatively stable height within a same comic story whereas it may vary in handwritten text, see Figure 2 and Figure 3.

The challenges related to speech text in comic books are the multiplicity of script, scale and orientation. Figure 1 illustrates the diversity of speech text in comic books. The text lines are quite short (space is limited in speech balloon) compared to other types of documents and, according to the style of the comics, there are also variations of writing style, case, spelling (word from dictionary or with voluntary spelling mistakes) and hyphenation. In order to cope with

Figure 2: Examples of typewritten text lines. The pixelization shows the low quality of certain images. Image credits: eBDtheque dataset.

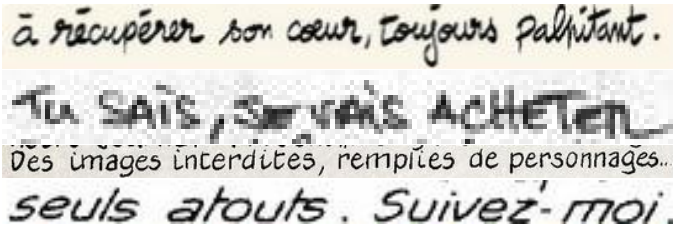


Figure 3: Examples of handwritten text lines. The pixelization shows the low quality of certain images. Image credits: eBDtheque dataset.

all these challenges we propose to analyze typewritten and handwritten text separately.

3. PROPOSED APPROACH

In this section, we briefly present a text localization algorithm we rely on and then review two well known OCR systems used for typewritten text recognition. Finally, we detail a specific training process for handwritten text recognition.

3.1 Text localization

In this study, we propose to rely on text localization to find speech text mainly from speech balloons. At this stage, any speech text extractor can be used but our result highly relies on its performance. In this paper, we use our algorithm that reaches 75.8% recall and 76.2% precision for text line localization (mainly speech text) on eBDtheque dataset [8].

3.2 Typewritten text recognition

Text recognition in comic book is one of the most complicated areas in comic document analysis. It is a complex process to recognize the text in comic book due to their various writing styles, different font size, complex background and degraded image quality. In this study, an attempt has been made to explore a recognition technique of comic text images. Two different OCR systems namely, Tesseract and ABBYY FineReader was considered for text recognition. Tesseract and ABBYY FineReader are two most popular OCR systems presently available. Tesseract OCR engine can be used in various operating systems. Tesseract, is considered to be one of the most accurate free open source OCR engines. Tesseract is not a complete featured OCR program. This open-source Tesseract OCR engine was originally developed at Hewlett-Packard between 1984 and 1994 [9] and

then improved by Google. ABBYY FineReader OCR engine can be used in a specific operating system. The pre-trained Tesseract OCR and ABBYY FineReader OCR have been considered for recognition in our experiment.

For recognition of the text images appeared in comic books, some pre-processing approaches have been considered in our experiment before feeding the text into OCR systems. Initially, the raw text images were converted into binarized text images. A global threshold is used to convert the gray images to binary images. After that, a smoothing technique was considered to remove the small noises. A resizing method with bicubic interpolation was also undertaken in the experimentation. For the recognition of comic text images, the Tesseract OCR engine and ABBYY FineReader OCR version have been adopted in this proposed study. Subsequently, the pre-processed text images were sent into the Tesseract and ABBYY FineReader to obtain the OCR output as an editable text.

3.3 Handwritten text recognition

As introduced in Section 1, handwritten text is very challenging for OCR systems. Their require a lot of annotated single letters of each font to train their optical model in order to be able to recognize text properly. In fact, it is not really feasible to annotate all letterer styles as they are continuously trying to be different from others (part of the identity of the comic book). Instead of annotating a huge amount of handwritten styles and build a generic handwritten OCR system, we propose to start by annotating a very small amount of data from a single letterer (person who writes text in speech balloons) and train the OCR on this specific letterer's style in order to reduce the Word Error Rate (WER) due to visual similarities (e.g. letter "i" from letterer A similar to letter "l" of letterer B). This approach has the advantages to require a small amount of training data, to reduce the complexity of the classification step of the OCR (less intra-class variability) and increase the OCR accuracy. Traditional OCR systems require annotated data at the level of letter which is really time consuming to annotate and inappropriate for cursive fonts [5]. Instead, we preferred to annotate text at the level of line using the efficient algorithm called OCRopus [1]. OCRopus is based on Long Short Term Memory neural networks (LSTM) that has proven its efficiency for handwritten text recognition [2].

The idea is to train the system on a minimum (few hundreds) of annotated text lines and then recognize all the other text lines from the same letterer (same writing style). This approach reduce image cropping and annotating time (groundtruthing) but may introduce false positives and false negative. False negative are not important because their will not decrease the quality of the ground truth, just ignore some text lines from the story. However, false positive may bias the ground truth so they must be manually removed by the annotator. Once the manual transcription have been done, we train the OCR on this writing style and then try to recognize all the other text lines from the same letterer.

4. PERFORMANCE EVALUATION

In this section we compare typewritten and handwritten speech text recognition quality using generic and specifically trained approaches. The idea is to highlight the limitations of standard pre-trained OCR systems at recognizing speech text in comics and measure the minimal human power which is required to get benefit from a font specifically trained OCR system and improve recognition rate.

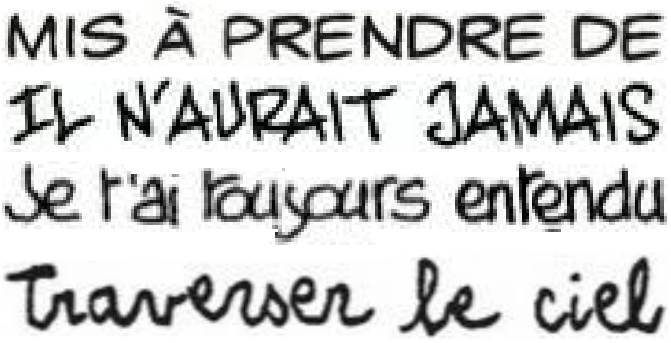


Figure 4: Example of text lines from album 1 to 4 (top to bottom). The pixelization shows the low quality of certain images. Image credits: Sequencity dataset.

Generic approach consists in measuring the ability of an OCR system that has been trained on a large amount of samples from a lot of fonts to recognize a single word. In a more specific approach, we measure the ability of an OCR system trained on a small amount of sample of a given font (writing style) to recognize text using this font and only this font. For both approaches, we rely on the Levenshtein Edit Distance algorithm for determining the recognition accuracy at character level [12]. We use the string similarity ratio which is given as a percentage between zero and one[1].

We evaluate the proposed method using text lines from two datasets: eBDtheque [4] and Sequencity[2] (online comics library) in order to show the performance of the presented recognition systems. We selected the eBDtheque dataset because it provides text transcription for all images (unlike Manga109). Sequencity dataset is related to a partnership we currently have with a company. The eBDtheque dataset was designed to be as representative as possible of the comics diversity, it includes few pages of diverse albums. It is composed by one hundred images which are composed by 4691 annotated 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%). From this dataset, we consider only the French comic books and webcomics in order to ease the comparisons with the second dataset which is only in French (3537 text lines).

We call the second dataset "Sequencity", it is a collection of more than 14 000 full albums that are for sale in French libraries or online. We selected four albums with different types of writing styles, Album 1 (ISBN: 9782369812500) is typewritten and contains only well separated uppercase letters. Album 2 (ISBN: 9782754815925) is also typewritten but with tilted uppercase letters and some touching letters. Album 3 (ISBN: 9782203073760) is handwritten in lowercase (except the first letters of sentences) and sometimes has touching letters. Album 4 (ISBN: 9782754808323) is handwritten with cursive text. An example of the writing style of each album is given Figure 4.

[1]https://rawgit.com/ztane/python-Levenshtein/master/docs/Levenshtein.html

[2]https://www.sequencity.com

Table 1: Average string similarity.

	Tesseract	FineReader	OCRopus
eBDtheque	0.53	0.56	0.82
Album 1	0.85	0.88	0.99
Album 2	0.38	0.21	0.96
Album 3	0.56	0.68	0.91
Album 4	0.28	0.24	0.80

Each albums contain between 130 and 500 pages, we annotated 500 text lines in each album from randomized pages. The groundtruthing was done by a French annotator following the process detailed Section 3.3. In this scenario, it takes approximately four hours to annotate one album (500 text lines) considering that the text extractor returns 50% of well segmented text line, the other 50% false positives are skipped by the annotator during the groundtruthing process.

The cropped text lines and associated groundtruth are available online[3]. However, the albums being not publicly available, we are not allowed to provide full images but the reader can get them images from the ISBN.

4.1 Results

In this section we compare the performance of three common OCR systems presented Section 3 (Tesseract, ABBY FineReader and OCRopus).

In this experiments, we did not re-train Tesseract and FineReader on the font (writing style) used in the image, instead, we used their pre-trained data for French language. However, we trained OCRopus on 50% of the annotated text lines and tested on the remaining 50%. It consist in 250 text lines for album 1 to 4 and 1768 text lines for eBDtheque dataset. The training process of OCRopus is quite simple, once the cropped text line images are assembled in a `folder` along with their transcription stored as text files, it can be performed by running a single command as below. The command outputs a model every thousand iterations. Detailed information are provided here[4].

```
ocropus-rtrain -o modelname folder/*.png
```

Note that text line images and corresponding transcriptions should be named as follows `image.png` and `image.gt.png` respectively.

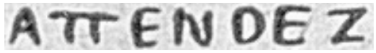
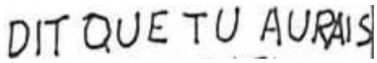
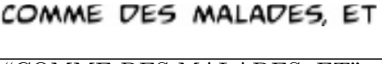
The percentage of training data and the number of iterations have been validated by experiments. The percentage of training data have been tested for 10%, 20%, 50% which represent 50, 100 and 250 text line images respectively (Table 4). The amount of training data is quite low because our aim was to measure the minimal human effort required for obtaining good recognition results.

For the number of iteration of the training process, we observed that 10 000 iterations were generally sufficient to train such system but we suggest to iterate over 50 000 iterations in order to be sure do not be affected by the lack of training issue.

Concerning the eBDtheque dataset in Table 1, the two generic OCR systems (Tesseract and ABBY FineReader) perform poorly compared to the specifically trained OCRopus but better than on Album 2 and 4. This is due to the diversity of fonts that they use during their training process to enforce their polyvalence. OCRopus gives promising results (82%) similar to its score on Album 4 (80%). However, Table 4 shows poorer results when it is trained with less than 50% of the dataset (50% training set, 50% testing set). Note that the standard deviation is between 20-30% for the three tested OCR systems. This poorness can be explained by the fact that images from this dataset contain various noises due to old printing techniques and low resolution scanning. Also because of the mixture of different writing styles that confuses the classifier of the OCR (Table 2).

For Album 1 to 4 in Table 1, their level of difficulty is clearly reflected by the performance of OCRopus. The latter decreases from 99% to 80% average accuracy in the same experiment condition (3% to 15% standard deviation respectively). OCRopus gives much better results than the two other OCR systems in general because the latter have not been trained on fonts similar enough to these writing styles including touching characters. Tesseract and FineReader perform in average better on Album 3 than the three other albums. It is probably because its writing style is non cursive and lowercase which is the most similar to typewritten

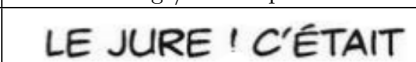
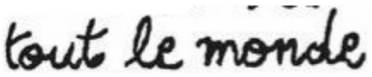
Table 2: Best result examples from eBDtheque dataset. Transcriptions are written between quotes below text line images for the ground truth (GT) and the three tested OCR systems, with corresponding string similarity value (Sim.).

OCR/im.	Image/transcription	Sim.
Image 1		
GT	“Pour faire”	
Tess.	“Pour faire ”	0.91
FineR.	“Pour fa ire”	0.91
OCRop.	“rour laire”	0.8
Image 2		
GT	“ATTENDEZ”	
Tess.	“A‘TTENOEZ ”	0.74
FineR.	“ATrew OEZ”	0.44
OCRop.	“A’TT’EN DE Z.”	0.76
Image 3		
GT	“DIT QUE TU AURAIS”	
Tess.	“DIT QUE TU gum ”	0.66
FineR.	“DlTQUÂčT(J AU$15”	0.48
OCRop.	“DIT OUE TU AU’AS”	0.84
Image 4		
GT	“COMME DES MALADES, ET”	
Tess.	“COMME ves MALADES, zr ”	0.72
FineR.	“COMMk MALAt7k6, kT”	0.55
OCRop.	“COMME DES MALADES ET”	0.98

[3]https://github.com/crigaud/publication/tree/master/2016/MANPU/toward_speech_text_recognition_for_comic_books

[4]http://www.danvk.org/2015/01/11/training-an-ocropus-ocr-model.html

Table 3: Best result examples for album 1 to 4 from Sequencity dataset. Transcriptions are written between quotes below text line images for the ground truth (GT) and the three tested OCR systems, with corresponding string similarity value (Sim.).

OCR/al.	Image/transcription	Sim.
Album 1		
GT	"LE JURE ! C'ÉTAIT"	
Tess.	"LE JURE ' C'êTAIT "	0.77
FineR.	"LE JURE ' C'ETAIT"	0.86
OCRop.	"LE JURE ! C'ÉTAIT"	1.0
Album 2		
GT	"UNE SANCTION MORALE"	
Tess.	"WE CANCTION W "	0.59
FineR.	"iwswcrwuitom"	0.00
OCRop.	"UNE SANCTION MORALE"	1.0
Album 3		
GT	"et parmi eux"	
Tess.	"'er' parrm eux. "	0.69
FineR.	"ef parmeox"	0.70
OCRop.	"et panmi eux"	0.92
Album 4		
GT	"tout le monde"	
Tess.	"Minnow "	0.29
FineR.	"t&xfc JlL7nanJ&"	0.28
OCRop.	"pout le monde"	0.92

font with which they have been trained for (Table 3).

5. CONCLUSIONS

This paper compares the performance of standard generic pre-trained OCR systems and to a specifically trained OCR system and measure the pros and cons. Both systems have been tested on a representative dataset as well as specific albums having different level of difficulties for an OCR system. The results analysis shows that pre-trained OCR systems perform fairly good on typewritten-like and lowercase writing styles. However, specifically trained OCR can recognize quite well all tested writing styles but requires some manual work for annotating the training samples. We measured the influence of the amount of training data to train such OCR system. This amount of training data is related to the number of writing styles to recognize and to its level of difficulty (e.g. uppercase only, mixed, cursive). In the future, we want to specifically train the OCR system on other writing styles and then combine them to build a speech text recognizer able to recognize the maximum of writing styles.

Table 4: Variation of the average string similarity according to the percentage of training data for OCRopus.

	Percentage of training data		
	10%	20%	50%
eBDtheque	0.40	0.70	0.82
Album 1	0.93	0.97	0.99
Album 2	0.91	0.95	0.96
Album 3	0.78	0.85	0.91
Album 4	0.48	0.57	0.80

Table 5: Variation of the accuracy according to the number of iteration of the training process.

	Number of iterations		
	10 000	50 000	100 000
eBDtheque	0.40	0.80	0.82
Album 1	0.97	0.99	0.99
Album 2	0.96	0.96	0.96
Album 3	0.90	0.91	0.91
Album 4	0.70	0.80	0.80

6. ACKNOWLEDGMENTS

This work was supported by the University of La Rochelle (France), the town of La Rochelle and the PIA-iiBD ("Programme d'Investissements d'Avenir"). We are grateful to all authors and publishers of comics images from eBDtheque and Sequencity dataset for allowing us to use and share their works.

7. REFERENCES

[1] T. M. Breuel. The ocropus open source ocr system. In *Proc. SPIE 6815, Document Recognition and Retrieval XV*, pages 68150F–15, 2008.

[2] T. M. Breuel, A. Ul-Hasan, M. A. Al-Azawi, and F. Shafait. High-performance ocr for printed english and fraktur using lstm networks. In *2013 12th International Conference on Document Analysis and Recognition*, pages 683–687. IEEE, 2013.

[3] M. R. Gaikwad and N. Pardeshi. Text extraction and recognition using median filter. *International Research Journal of Engineering and Technology*, 3(1):717–721, 2016.

[4] 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)*, pages 1145–1149, Washington DC, 2013.

[5] M. Heliński, M. Kmieciak, and T. Parkoła. Report on the comparison of tesseract and abbyy finereader ocr engines. 2012.

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

[7] S. Ranjini and M.Sundaresan. Extraction and recognition of text from digital english comic image using median filter. *International Journal on Computer Science and Engineering (IJCSE)*, 5(4):238–244, April 2013.

[8] C. Rigaud, D. Karatzas, J. Van de Weijer, J.-C. Burie, and J.-M. Ogier. Automatic text localisation in scanned comic books. In *9th International Conference on Computer Vision Theory and Applications*, 2013.

[9] R. Smith. An overview of the tesseract ocr engine. In *Proc. Ninth Int. Conference on Document Analysis and Recognition (ICDAR)*, pages 629–633, 2007.

[10] P. Srikanta, B. Jean Christophe, P. Umapada, and O. Jean Marc. Line-wise comic text identification: A support vector machine-based approach. In *Proceedings of International Joint Conference on Neural Network*, pages 1996–2000, Vancouver, Canada, 2016.

[11] C. Y. Su, R. I. Chang, and J. C. Liu. Recognizing text elements for svg comic compression and its novel applications. In *2011 International Conference on Document Analysis and Recognition*, pages 1329–1333, Sept 2011.

[12] S. Tanner, T. Muñoz, and P. H. Ros. Measuring mass text digitization quality and usefulness. *D-lib Magazine*, 15(7/8):1082–9873, 2009.