VIETNAMESE TEXT EXTRACTION FROM BOOK COVERS

Phan Thi Thanh Nga[a], Nguyen Thi Huyen Trang[a], Nguyen Van Phuc[b], Thai Duy Quy[c], Vo Phuong Binh[a*]

[a]The Faculty of Information Technology, Dalat University, Lamdong, Vietnam
[b]The Devsoft Company, Hochiminh City, Vietnam
[c]The Research Management and International Cooperation Department, Dalat University, Lamdong, Vietnam

Article history
Received: January 09th, 2017 | Received in revised form: April 19th, 2017
Accepted: May 11th, 2017

Abstract

Automatic information extraction from images reduces the cost, human interference, and timely processing. Converting printed book covers to readable text for later automation process would be useful for a wide range of users such as librarians, bookshop keepers, and individual users. In this paper, we present a novel method for the Vietnamese text extraction from images of scanned book covers. The proposed system accepts the book covers snapshot, filters the input image for an enhancement of quality, locates the regions with text, then utilizes the optical character recognizer (OCR) to extract the text. The last step is to filter the extracted text in accompany with at dictionary to achieve the final text result. Carrying out the experiments with the proposed system using our dataset delivered encouraging experimental results.

Keywords: Book cover; OCR (Optical Character Recognition); Text information extraction; Vietnamese text detection.

1. INTRODUCTION

Before the existence of computers, books were often purely manually categorized using library classification systems, commonly known as DDC (Dewey Decimal Classification), LCC (Library of Congress Classification), CC (Colon classification), UDC (Universal Decimal Classification). However, these considerably disciplinary systems are preferable by librarians when people tend not to follow the hierarchy structure in their daily archive. Thanks to the help of technologies, books can be recognized by covers. Information about book titles, authors, or publishers would be a valuable input

* Corresponding author: Email: binhvp@dlu.edu.vn

for tracking systems or augmented applications. Existing systems use feature-based image matching to obtain accurate book covers recognition (Chen, Tsai, Vedantham, Grzeszczuk, & Girod, 2009; Matsushita, Iwai, & Sato, 2011). Databases of book covers in the mentioned systems are limited number of books written in English or books supported by Amazon API.

Instead of being much dependent on book covers image databases, our approach is to extract the text directly from the book covers. Our target is the text in Vietnamese of the book cover because this piece of information is really helpful in retrieving further information of a book using simple text queries. Acknowledging the sophisticated scene of book covers, we investigated the areas of OCR and image processing to find robust methods to detect and retrieve text from the scene backgrounds. The proposed system accepts scanned book covers images. The images are then reprocessed via different filters by either de-skewing skewed images, denoising noised images, binarizing images, removing complex backgrounds, or detecting text regions in images. The filtered images are then provided to the OCR engine; the open source OCR engine Tesseract which is utilized due to its Apache license would allow us to add more value to the provided code. Although commercial OCR engines such as VnDOCR and ABBYY provided the solution to work with Vietnamese text, they are limited to serif text on simple backgrounds. The proposed system aims to start with those steps extracting the text from the book covers and the extracted text would be refined to be the input of other tracking or augmented systems.

In this paper, we first explore the current state of text recognition approaches and investigate areas where improvements could be made. Section 3 introduces the proposed system architecture and Section 4 provides more information on detailed implementations and discusses the advantages and shortcomings of the proposed approach. In the last Section, we provide the system demonstration by the results.

2. LITERATURE REVIEW

Text recognition from sophisticated backgrounds has been an interesting topic due to its vast benefits in many systems (Gatos & Pratikakis, 2005; Too & Prabhakar, 2016;

Sobottka, Bunke, & Kronenberg, 1999; Yadav, 2015; Zhong, Karu, & Jain, 1995; Zhu, Yao, & Bai, 2014). Texts with different sizes and fonts along with complicated background are supposed to be parts of book covers. Book cover recognition, as the part of scene text recognition, inherited both the problems and solutions researched in the area of scene text recognition. Solution for detecting text region in images with background was proposed early in the problem of book and journal covers recognition (Sobottka et al., 1999).

In the pre-processing step, there is need to filter images. ImageMagick library, provided under the Apache 2.0 license at the URL http://www.imagemagick.org/script/index.php, comes up with rich yet efficient functionalities to work with images. Using ImageMagic, we may rotate the image by 90 degrees, un-rotate images limited to less than five degrees, crop, convert grayscale, enhance the image quality, or smooth texts.

The problem of skewed images is found in OCR projects. Since 1989, the Hough transform has been proposed to detect the skewed angle of the images and was proven to be a efficient way to detect skewed text (Rosner, Boiangiu, Zaharescu, & Bucur, 2014; Srihari & Govindaraju, 1989). The Hough transform technique detects the presence of a parametrically representable group of points in an image, such as a straight line or a circle through a mapping to a parameter space. Hough transform can work with various types of lines but we focus on how it works with straight lines. The idea behind this is that there are numbers of lines passing through a given dot. To obtain the lines pass through the largest number of dots, we calculate the highest value of dots those votes the lines pass. For text, we choose the lines that satisfy that most of the dots above them are in black while the dots below them are in white. This feature of the chosen lines ensures that the lines are the bottom lines of the text.

OCR in non-English languages might end up with several problems. Vietnamese is a language with tones and single syllables. We were not successful in finding any related studies in Vietnamese scene text extraction, but there are some efforts in other

scripts language such as Bangla (Chowdhury, 2016; Hasnat, Chowdhury, & Khan, 2009a, 2009b).

3. VIETNAMESE TEXT EXTRACTION FROM BOOK COVERS (VTEB)

The three main stages of the VTEB to detect Vietnamese scripts into text (OCR pre-processing, OCR, and OCR Posprocessing) are illustrated in Figure 1.

In OCR processing, the main purpose is to clean the input image. Different image filters would be applied in order to remove the noise, and/or color. The image would also be de-skewed in case there are any skewness detect the text region once successfully found would help the OCR stage. This step is done under the umbrella of the ImageMagick library. Although most of the pre-processing step could be fulfilled by the library, we apply the Hough transform for a better de-skew result. We also apply the Sobel filter for the text region detection. We first take a deep look in the Hough transform and apply it into our system as follows:

(1) Find all of desired feature points in the image

(2) Foreach feature point

(3) Foreach possibility *i* in the accumulator that pass through the feature point Increment that position in the accumulator

(4) Find the local maxima in the accumulator

(5) If desired, map each maxima in the accumulator

The pseudo algorithm would be as follows:

(1) Create the two-dimensional matrix with all initialized value set as 0.
(2) For y=0 to height
(3) For x=0 to width
(4) If (x, y) is black
(5) For alpha=90 to -90
(6) d=abs(y*cos(alpha)-x*sin(alpha))
(7) Find the value of (abs(alpha*5), d) in the matrix and increase it by 1
(8) Increase alpha by delta

(9) Scan the matrix and find the top 20 values

(10) Calculate the average of the top 20 values

(11) De-skew the image by the average value in the previous step.

The algorithm using for skewness text detection would use the black point in the image to start finding the text. We then find the lines crossing the point and each point can vote for the lines passing through it. The line with the highest vote would be chosen. The algorithm is time consuming especially when working with large sizes, and high quality images. We reduce the scanned angle to [-16,16] as we found books should not be skewed more than that. Therefore, the algorithm could work with any scanned angle when the user may configure these values. The angle once detected will be used to rotate the image back and eliminate the skewness.

The OCR stage help parse and recognize the text from this the image provided by the previous step. In this stage the OCR is consultAmong the OCR engines provided, we chose Tesseract OCR because of its openess and richness. To work with Tesseract, we first conducted the learning process feeding different font faces and font sizes. These training process helped Tesseract to understand the language well. The trained data would then be feeded into Tesseract so that the engine would work more efficiently with the provided data.

The final OCR post-processing was then used to clean text and fix the error of the output OCR. We did not fully implement this step since we used the dictionary embedded in the Tesseract engine itself. The basic behind this step is that the text would be cleaned using the dictionary. Uncommon words would have negative points when common words would obtain the positive points.

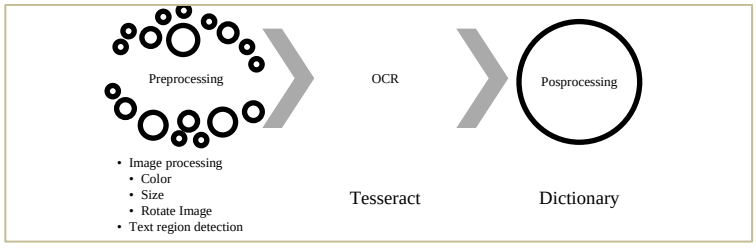
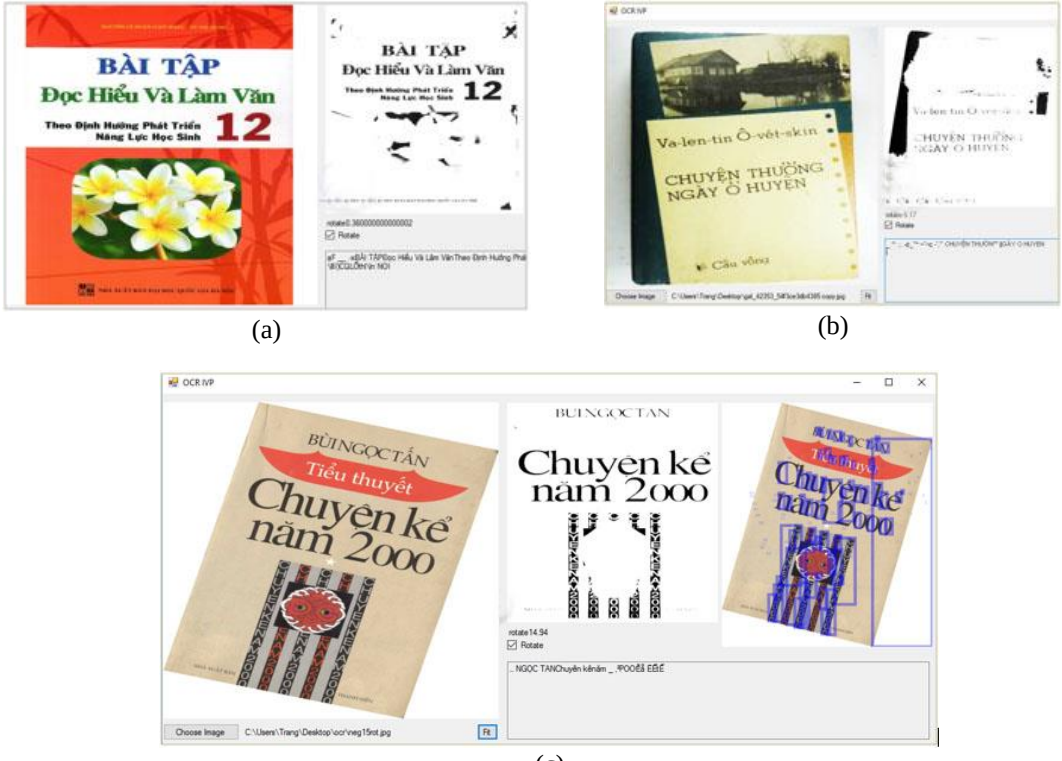


Figure 1. Approach of the proposed system

The proposed system is built in .NET framework utilizing the above-mentioned steps. The proposed system was organized as the libraries and modules could be added when necessary. The first intention was of a general framework.

We implemented the system using databases of textbooks available in Vietnam. The textbooks are captured images with rotation, background, the mix of different font sizes and font types. These images are retrieved manually by taking photos from book covers or online results.

Use cases: By walking through a scenario that spans the detection process, we wish to highlight the major benefits of our solution. We now demonstrate the VTEB system with the use cases: (1) Simple background image; (2) Skewed image; (3) Complex image background. Figure 2 illustrates the result of these use cases processed by our system.



(a) (b)

(c)

Figure 2. Illustration for case studies

Note: (a) Scanned book cover with simple background; (b) Scanned book cover being skewed (c) Skewed scanned book cover with sophisticated background.

Simple background colored book covers: The image would then be processed to remove the noise and color. It would then be passed to the OCR engine and the text value would then be displayed in the result box. As shown in Figure 2a, VTEB could read 75.56% of the book cover except the yellow line on top due to the low contrast in that area. The percentage of accuracy is calculated according to the rate of right words recognized over total words counted.

Skewed images: Scanned images must first be filtered to eliminate the color, noise, and the simple colored background. When the scanned book cover is skewed, VTEB checks the image to decide the skewed angle of the image and automatically rotates the image back to obtain the unscrewed image. This step is mandatorily set for every image provided to the system though the user has the power to discard this step in the system. The filtered image would be fed to the OCR engine and the text would be extract the extracted value would later be cleaned using the dictionary and the final value will be displayed on the result box. The illustration shows the noise in result text and this is due to the text written on the illustration of the book cover itself. As shown in Figure 2b, the text in the book cover is recognized with an accuracy of 78.3%.

Complex image backgrounds: When the image comes up with a sophisticated background, VTEB would try to detect the text area. These text areas will then be provided to the OCR engine. Similar to the previous case, scanned images must first be filtered to eliminate the color and/or noise. The filtered image would be fed to the OCR engine and the text would be extracted. The extracted value would later be cleaned using the dictionary and the final value will be displayed on the result box. As shown in Figure 2c, we reached an accuracy of 82%.

The purpose of the proposed system is to obtain clean texts from scanned book cover images. Our system accepts scanned images at the resolution of 300dpi and the size is 300px each dimension for the smallest. The image will then be added to the pre-processing phase which would either de-skew, de-noise, enhance, crop, or remove background of the image. This filtered image would then be provided to the OCR engine-Tesseract in this case. The OCR engine, using the trained data language, helps us to detect

the text from an input image. The retrieved text should then be refined by going through a post-processing step using the dictionary. The final result would then be obtained and displayed.

By implementing VTEB, we the set up for our system came up with these benefits: accepting Vietnamese fonts in both serif and san-serif, eschewing the images, filtering the background, and detecting text area automatically. The advantage of our system is that it can work with Vietnamese text typing in both serif and san-serif fonts because we applied the training set of tesseracts in both of these fonts. Skewed images could be recognized and processed automatically. The result of the skewness detection and de-skewing in our application is described in Table 1. The background is filtered using ImageMagik and text areas are detected.

Table 1. Result of de-skewing process using Hough transform

Right angle	Detected angle
0.2	0.36
15	14.98
5	5.02
-14	-13.95

Besides, there are limitations as follows: The resolution of the input might have some effects on the output result, the mix of different font sizes and font types might lead to incorrect results. The provided book cover scanned image at the resolution of 300dpi and the size is 300px each dimension for the smallest. Low contrast text-background could not be processed. As shown in Figure 2c, the mixed font size might lead to wrong results: The number 2000 is recognized as 000 because the number 2 is not in the same base line as the other texts recognized. The rotated angle, if there is any, should not be larger than 16 degrees. The configuration could help VTEB work with larger degrees but this might lead to a slow recognition process. Text areas are not fully recognized. This function needs future modification. The post-processing step is not yet implemented which may lead to the unclean result. We also had a problem when no databases open for Vietnamese book covers are available to test. We left this for our future work.

4. CONCLUSION

By implementing VTEB from scratch, we pointed out that this project can be considered as an initial trial of the book cover recognition systems. We provided the workflow and our implementation to retrieve clean text out of book covers images. The Hough transform was implemented to de-skew the skewed images. We also utilized different filters in order to reduce the noise, remove the background, and extract the text regions. Thanks to Imagemaker for the enormous help in cleaning the source image with basic filters. The use of Tesseract engine helps in OCR the text from input images. In the prototype, we have loosely implemented the steps of image processing, OCR, and image post-processing. However, each of these steps could be changed independently for the appropriate result.

The response time of the Hough transform algorithm is considerably slow in case we set the large value to the angle. By assuming that most skewed images receive the value less than 15 degrees, we decrease the complexity of the algorithm by limiting the angles in [-16, 16] degrees and the step size is set to 0.2.

The paper provides the solid foundation for text extraction from book covers and our implementation would be integrated into future research for excessive functionalities. We did not rush to implement the fully functional system but rather ensure that we obtain the right approach at first. And this leads to the worked but yet simple user interface. However, the limitations of the systems need to be addressed and solved in the upcoming version of the system.

We are planning to further implement the system on mobile devices. This requires us to work more on the algorithm complexity. We also need to conduct a larger range of test databases and a detailed statistical test on our system.

REFERENCES

Chen, D. M., Tsai, S. S., Vedantham, R., Grzeszczuk, R., & Girod, B. (2009). *Streaming mobile augmented reality on mobile phones.* Paper presented at The IEEE International Symposium on Mixed and Augmented Reality, USA.

Chowdhury, A. (2016). Bangla character recognition for Android devices. *International Journal of Computer Applications, 136*(11), 13-19.

Gatos, B., & Pratikakis, I. (2005). *Text detection in indoor/outdoor scene images*. Paper presented at The First Workshop of Camera-Based Document Analysis and Recognition, Spain.

Hasnat, M. A., Chowdhury, M. R., & Khan, M. (2009a). *An open source Tesseract based optical character recognizer for Bangla script*. Paper presented at The International Conference on Document Analysis and Recognition, Spain.

Hasnat, M. A., Chowdhury, M. R., & Khan, M. (2009b). *Integrating Bangla script recognition support in Tesseract OCR*. Paper presented at The Conference on Language and Technology, Spain.

Matsushita, K., Iwai, D., & Sato, K. (2011). *Interactive bookshelf surface for in situ book searching and storing support*. Paper presented at The 2nd Augmented Human International Conference on - AH '11, Japan.

Rosner, D., Boiangiu, C., Zaharescu, M., & Bucur, I. (2014). *Image skew detection: A comprehensive study*. Paper presented at The Third International Workshop on Cyber Physical Systems, Romania.

Sobottka, K., Bunke, H., & Kronenberg, H. (1999). *Identification of text on colored book and journal covers*. Paper presented at the Fifth International Conference on Document Analysis and Recognition, Spain.

Srihari, S. N., & Govindaraju, V. (1989). Analysis of textual images using the Hough transform. *Machine Vision and Applications*, *2*(3), 141-153.

Too, K. B., & Prabhakar, C. J. (2016). Extraction of scene text information from video. *International Journal of Image, Graphics and Signal Processing*, *8*(1), 15-26.

Yadav, N. (2015). *Algorithm for automatic text retrieval from images of book covers.* (Master Thesis), Thapar University, India.

Zhong, Y., Karu, K., & Jain, A. K. (1995). Locating text in complex color images. *Pattern Recognition, 28*(10), 1523-1535.

Zhu, Y., Yao, C., & Bai, X. (2016). Scene text detection and recognition: Recent advances and future trends. *Frontiers of Computer Science, 10*(1), 19-36.

NHẬN DẠNG BÌA SÁCH TIẾNG VIỆT CHO ỨNG DỤNG QUẢN LÝ SÁCH

Phan Thị Thanh Nga[a], Nguyễn Thị Huyền Trang[a], Nguyễn Văn Phúc[b], Thái Duy Quý[c], Võ Phương Bình[a*]

[a]Khoa Công nghệ Thông tin, Trường Đại học Đà Lạt, Lâm Đồng, Việt Nam
[b]Công ty Devsoft, Thành phố Hồ Chí Minh, Việt Nam
[c]Phòng Quản lý Khoa học và Hợp tác Quốc tế, Trường Đại học Đà Lạt, Lâm Đồng, Việt Nam
[*]Tác giả liên hệ: Email: binhvp@dlu.edu.vn

Lịch sử bài báo
Nhận ngày 09 tháng 01 năm 2017 | Chỉnh sửa ngày 19 tháng 04 năm 2017
Chấp nhận đăng ngày 11 tháng 05 năm 2017

Tóm tắt

Nhận dạng văn bản từ hình ảnh giúp giảm công sức, chi phí và thời gian xử lý. Việc số hóa thông tin sách một cách tự động bằng cách nhận dạng bìa sách giúp ích rất nhiều cho những người làm việc trực tiếp đến lưu trữ và phân loại sách như thủ thư, nhân viên nhà sách và kể cả những người dùng cá nhân chỉ muốn quản lý một thư viện cá nhân tại nhà. Trong bài báo này, chúng tôi đề xuất phương pháp nhận dạng văn bản tiếng Việt từ ảnh bìa sách. Hệ thống xử lý ảnh bìa sách ở đầu vào, chỉnh ảnh để đạt được ảnh phù hợp cho quá trình nhận dạng, định vị các vùng chứa văn bản, sau đó áp dụng kỹ thuật nhận dạng ký tự quang học (OCR) nhằm thu được văn bản chứa trong ảnh, bước cuối cùng chúng tôi lọc nội dung rút trích ở bước trên và sử dụng từ điển để nâng cao độ chính xác của văn bản nhận diện được. Chúng tôi tiến hành kiểm tra chương trình và nhận được kết quả khả quan cho bộ dữ liệu bìa sách được đưa vào thử nghiệm.

Từ khóa: Bìa sách; Nhận dạng tiếng Việt; Nhận dạng văn bản.