Text Recognition and Translation of Multi-Oriented, Multi-Language and Curved Text in Natural Scene Images

Cristina Vassallo

June 2015

Submitted to the MCAST Institute of Information and Communication Technology in partial fulfilment of requirements for the Bachelor of Science (Honours) in Software Development.
Abstract

This study is about text detection and recognition in natural scene images. The main focus is on the detection, recognition and eventually, translation, of multi-oriented, multi-language and curvilinear text in such images. The study attempts to provide a solution that can detect and recognise such text since current leading mobile applications such as Word Lens and Google Goggles do not support such text for translation.

There are many algorithms available that can detect and recognise text, but very few consider text which is multi-oriented, multi-script or curvilinear. Text detection can be carried out using various methods including region-based and texture-base methods. Furthermore, algorithms for multi-oriented text detection are further divided into non-headline and headline based methods. Three different solutions were considered in this study, one being an algorithm developed specifically for maps in which text usually has various orientations, curvatures and sizes. Another option was a framework that performed detection and recognition simultaneously. The third option was a combination of two algorithms, one for detection and one for aligning curved and multi-oriented text.

The research carried out consisted of implementing the final option and integrating these two algorithms to achieve a system that could detect and recognize multi-oriented, multi-script and curvilinear text. Tests were carried out using the proposed system by using different data sets that are publicly available for testing such systems. Results were recorded according to the criteria of time, precision and recall. In addition, the proposed system was compared to two leading applications: Word Lens and Google Goggles.

The result of this study was that the resulting system was not capable of performing detection and recognition of all kinds of images and text. On the other hand, it was able to give a solution for issues that other leading applications face, such as the detection of multi-oriented and multi-script text. Results can be affected by various aspects, including the libraries and languages used to implement an algorithm. In addition, the device used and its processing power are highly related to the performance and probably also to the precision of a system. Other factors that need to be considered when implementing such a system are the datasets that will be used for testing and the experience and knowledge one has in the area.
Introduction

Smartphones have become indispensable and can be used for many purposes, especially whilst travelling. Lately, applications such as Word Lens have become very popular with travellers who would like to translate text in a foreign language into their native language. Such apps allow one to take a picture of the text or just point their mobile camera at it and the original text is detected and translated it to the source language. However, a common problem for these apps is that text can come in various sizes, colours, fonts, orientations and curvatures and due to these factors, it may not always be read correctly. The main objective of this study is to offer a solution which can detect curvilinear text, for instance in an arc-form, or text which is not horizontally aligned but has a different orientation, for instance vertically aligned or slanting text, to be able to improve the text detection capabilities of applications like Word Lens. The algorithm used should also be able to detect different scripts so as not to limit the use of the solution for languages using only the Latin script, such as English, Italian and Spanish.

When researching my dissertation topic, I was mainly interested in mobile applications because of the shift from web apps to mobile apps and thus the increased need for software developers specialised in mobile app development. Having encountered the app Word Lens, my interest in the area of text detection, and recognition in particular, was piqued. This area is still relatively new and further research is expected, especially with regards to detection and recognition of multi-oriented and curved text in natural scene images. The kind of apps mentioned previously can be very useful whilst travelling and if text detection and recognition can be improved further, such applications would eventually be able to offer better translation service to their users, specifically when travelling to countries where instant and on the spot translating is imperative.

As part of the analytical process, two existing algorithms were integrated to perform text detection and recognition of text having different orientations or curvatures. The first algorithm is a text detection algorithm that detects text based on perceptual organisation, offering support for the text detection of multi-oriented and curved text strings in different scripts. This algorithm has achieved good results when compared to other algorithms that perform the same function, placing second with regards to precision but obtaining a much better performance in relation to the time taken to perform the text detection. This algorithm
was naturally chosen as it offers a balanced implementation, taking into consideration both precision and performance. The second algorithm is an algorithm that given the detected text as a binary image, horizontally aligns the characters no matter the orientation or curvature of the original text. The latter is used to improve the Optical Character Recognition (OCR) results since most common OCR software offers limited support for multi-oriented or curved text. This method has been proven to improve significantly the text recognition results by aligning the text horizontally and maintaining the inter-character and inter-word spacing of the words to obtain meaningful translation. The result obtained after using these algorithms is then processed by an OCR for text recognition and can eventually be translated.

These two algorithms were chosen because their integration can help improve text detection and recognition of multi-oriented and curved text on natural scene images. As stand alones they offer limited features, but when combined they can offer a much more complete system for end-to-end text recognition. The integration of these two algorithms can help compare results of text recognition before, and after, alignment of characters as the orientation and layout of the text lines play an important role in performing OCR.
Conclusion

Text Detection

The time taken to perform text detection was not ideal. Various factors contributed to this result. The detection algorithm does not make use of the GPU, making the process much slower. This is a point to consider when choosing an algorithm as a GPU is not available in all devices, thus it will limit its use to certain kinds of devices. Mobile devices, for instance, have limited processing power when compared to PCs and thus, such an algorithm would not be adequate.

The algorithm chosen for text detection made use of the OpenCV\(^1\) library for functions related to image processing and was implemented in C++. The choice of the libraries and languages used to develop an algorithm can affect the performance as well. The use of Emgu CV\(^2\) for implementing this algorithm in C# could have had a serious impact on its performance, since by using a wrapper to OpenCV there is an extra layer to go through that could have slowed down the algorithm. In addition, the use of C# might not have been ideal for this kind of application. C# is generally as fast, or even faster, than C++ but in relation to graphics and image processing, the use of pointers in C++ make it a better option performance-wise.

The precision and recall rates achieved when compared to those for the algorithm by Gomez and Karatzas (2013) are slightly below their values. The algorithm used for text detection was implemented in C# and Emgu CV as discussed, and this could have affected not just the performance but also result precision. The fact that the GPU was not used, limiting the memory and resources available for processing, could have meant a lack in precision during the detection phase.

The use of the edge-enhanced MSER technique to increase the number of words detected and therefore, decrease the number of false negatives could have increased the false positives, resulting in a low recall rate for the algorithm. Moreover, the grouping of the words into phrases and eventually into sentences might not have been implemented the right way. Implementing an algorithm that can be quite complicated, such as that proposed by Kasar

\(^1\) [http://opencv.org/](http://opencv.org/)
\(^2\) [http://www.emgu.com/](http://www.emgu.com/)
and Ramakrishnan (2013), nearly from scratch without having at least some experience in the area might not be ideal and thus lead to erroneous results. In addition, the code to finally group the detected phrases into text lines was implemented completely from scratch without any reference to the best way for achieving text line grouping. Another important aspect for achieving better results is the filtering out of connected components (CCs), which are definitely not part of the text in the image, such as elements that are too large when compared to the text characters in the image.

**Text Alignment**

The text alignment algorithm is quite fast although the same aspects mentioned for text detection could allow for improvement in this algorithm as well. This algorithm was used improve text recognition for multi-oriented and curved text but the results do not show an improvement due to the fact that the Optical Character Recognition (OCR) software did not provide correct results. Another factor that could have contributed to this result is that the datasets used are not specific for curved text and thus the alignment algorithm could have improved text recognition in cases where the text was curved or slanted but might have affected badly recognition of the horizontally aligned text. Unfortunately the dataset used to test the original implementation of this algorithm was not available for testing by the public. Therefore having a dataset specific to the system could help to better assess it. Furthermore, the sample images shown in Kasar and Ramakrishnan’s (2013) study contain only Latin scripts, which makes it quite safe to assume that the images in their dataset contained only text from the Latin alphabet and that the Chinese characters in the MSRA-TD500 dataset might have had an impact on the result. In conclusion, this algorithm proved to be useful for aligning curved or multi-oriented text but proved inadequate for scripts other than Latin-based letters, in some cases incorrectly aligning text which was already horizontally aligned.

**Existing Applications**

The resulting application proved capable of detecting vertical and slanted text in different language scripts whereas other applications are not able to do this. On the other hand, the result for curved text was expectedly good, as Google Goggles proved to be better for cases considered for testing. The application was quite robust with regards to mixed scripts in one image. The other applications tested could only detect a few words or just words in English.
Even though the application is not ideal for all kinds of images as the results show, in some aspects of text detection and recognition, the proposed system proved to be better than available applications. The techniques presented in this study can be applied to already existing applications to target multi-script and multi-oriented text and improve their overall detection and recognition results.

Future Work

This study contributes to a very larger area of study since text detection and recognition is quite vast. Following this study there are a number of possible avenues that can be further investigated, including:

- The use of text recognition during text alignment such that upside down words can be identified and rotated accordingly.
- Improved filtering of CCs which are definitely not part of the text in the image.
- A text alignment algorithm that supports other scripts and languages such as Chinese.
- OCR software that can simultaneously process different languages within the same image.

Conclusion

This study has proposed a system consisting of a perceptual organisation text detection algorithm, an edge-enhanced MSER detection algorithm and a text alignment algorithm that align horizontally multi-oriented and curved text so as to improve recognition results. The system includes an OCR and translation software to be able to perform text recognition and translation of the detected and aligned text.

The proposed solution was tested using the MSRA-TD500 and HUST-TR400 datasets. The analysis included time taken by algorithm to perform detection and recognition, precision, recall and f-score. The system was also compared to leading application Word Lens and Google Goggles.