A PRECISE SKEW ESTIMATION ALGORITHM FOR DOCUMENT IMAGES USING KNN CLUSTERING AND FOURIER TRANSFORM

Jonathan Fabrizio

LRDE-EPITA
14-16, rue Voltaire,
94276 Le Kremlin-Bicètre CEDEX France
jonathan.fabrizio@lrde.epita.fr

ABSTRACT

In this article, we propose a simple and precise skew estimation algorithm for binarized document images. The estimation is performed in the frequency domain. To get a precise result, the Fourier transform is not applied to the document itself but the document is preprocessed: all regions of the document are clustered using a KNN and contours of grouped regions are smoothed using the convex hull to form more regular shapes, with better orientation. No assumption has been made concerning the nature or the content of the document. This method has been shown to be very accurate and was ranked first at the DISEC'13 contest, during the ICDAR competitions.

Index Terms— Skew estimation, KNN, Fourier transform.

1. INTRODUCTION

Many document algorithm are affected by skew distortion. In order to get the best possible accuracy for such processes, it is essential to have a precise skew estimation technique. There already exist several approaches to estimate the skew of a document. However research in skew estimation algorithm is still a research area [1].

An overview of different investigated strategies can be found in [2] and more recently in [3].

A lot of algorithms share the same approaches. The usage of Hough transform has extensively been tested [4, 5]. Many different variations have been tested such as applying the Hough transform to the centroid of regions, boundaries, combining with RLE, etc. Another investigated strategy is the projection profile [6, 7]. These skew estimation methods rely on the horizontal or vertical accumulation of pixels. Clustering methods have also already been tested [8, 9]. Most of the times they consist in detecting aligned elements to form chains and then evaluate the slope of these chains. Morphological operators can also be used [10]. The usage of mathematical morphology has the advantage to manage grayscale images easily. Finally, the freqeuncy domain with the Fourier transform or DCT [11, 12, 13] has sometimes been used.

Most of the times, the projection profile or the Hough transform-based method can not be applied to any type of documents. For example they might fail on comics. Methods based on clustering or methods that estimate skew in frequency domain might have more chances to be generic.

The need to increase the precision of skew estimation led us to develop this method. Our goal is to make as few assumptions as possible about the nature or the layout of the document. This is why our method uses the magnitude spectrum of a Fourier transform to determine the orientation of the document image. The orientation of main frequencies in this representation is directly linked with the orientation of image elements. A rotation in the spatial domain lead to the same rotation in the magnitude spectrum. However using directly the frequency representation of the document image brings coarse results. By preprocessing the document image with a clustering algorithm the result is enhanced.

Our method does not need any assumption concerning the document style (newspapers, comics...) or its content (alphabet...). It has proven to be very precise and robust. The method works on binarized documents or, at least requires a segmented or labeled document.

The article is divided as follows. In section 2 we explain the algorithm step by step and illustrate it. Then we show the precision of the algorithm mainly by analyzing DISEC'13 contest results [1] in section 3, which is followed by the conclusion in section 4.

2. ALGORITHM

Our goal is to estimate the skew of the document image. As seen before, several strategies to estimate the skew of an image document already exist. Our algorithm takes advantage of the fact that a rotation in the spatial domain leads to a rotation in the magnitude spectrum. However computing the Fourier transform on the original image and then analyzing the frequencies to estimate the rotation angle is not easy and

IEEE ICIP 2014
DOI: 10.1109/ICIP.2014.7025523
http://ieeexplore.ieee.org/document/7025523/

the result will be particularly imprecise.

Figure 1 (a)-(b) and Figure 2 (a)-(b) show common document images and their Fourier transforms. Results of Fourier transforms do not have enough contrast: the correct direction is not sufficiently highlighted. The document is then preprocessed and all regions of the document are clustered using a KNN. During the next step, the Fourier transform is applied to the image of the outlines of the convex hulls of the clustered regions. In that way, in the frequency domain, the orientation is easier to be detected.

Document image preprocessing: KNN Clustering The document image has to be processed to enhance essential frequencies. The amplitude of other frequencies must be reduced. Tortuosity of elements (characters...) must be smoothed to reduce noise in the frequency domain. Our target here is then to simplify the document image, as much as possible. In order to simplify the outline of shapes in the document and to keep only the outline in the correct direction (or make them come into being), the regions of the image are clustered many times using a KNN clustering with a simple strategy.

The goal is to combine similar regions that are close to each other. For all the regions, we look for k nearest neighbors that have a comparable thickness and size and whose distance is proportionnal to the elongation of the region. Then to form groups, a region I is linked with a region J if and only if I belongs to k nearest neighbors of J, and J belongs to k nearest neighbors of I. At the end of the process we get multiple sets of regions. To get a usable and smooth result we keep only convex hull boundaries/contours of each set. In practice we apply this process twice with $k = 6$ and $k = 9$ and keep the union of the two results.

Figure 1 (c) and Figure 2 (c) show the union of results of clustering. We only keep the outlines of convex hulls of clustered regions. Figure 1 (d) and Figure 2 (d) are Fourier transforms of Figure 1 (c) and Figure 2 (c) respectively. Interesting frequency has been enhanced while other around has been soften. Searched rotation angle is now clearly brought out.

Skew angle estimation from frequency domain Once the document is preprocessed, the estimation is performed in the frequency domain (Figure 1 (d) and Figure 2 (d)). The main orientation of the document is now clearly visible in this frequency representation. Only its precise extraction remains now. Multiple approaches can be used to extract this orientation and we have tested some of them (inertial axis...). In our tests, the approach that leads to the most precise result is the follower. We define a cross C (that is aligned with the horizontal and the vertical axis). We note $R(I, \alpha)$ the rotation of image I according to angle α. We take I the frequency representation of our preprocessed document and then compute:

$$t(\alpha) = \sum R(I, \alpha) * C \quad (1)$$

The searched angle is the α that leads to the higher value for $t(\alpha)$. There are certainly other strategies that can improve the precision of the angle extraction however we have already reached a good precision.

Possible alternatives of the algorithm From this method we can derive various improvements. First we select experimentally parameter k with $k = 6$ and $k = 9$. However there might be refinements here to select correct values of k. Small values of k will favor grouping of letters into words. Higher values of k will form sentences and paragraphs.

The second improvement consists in computing clusterings of regions in the image, but also clusterings of regions in the negative image and take the union of the two results. In our tests the result is less precise but more robust to the diversity of documents. But we do not conduct enough tests to definitely conclude. There is also probably a strategy to select correct values k for the image and correct values k for the negative image. Using the negative image seems to be particularly well adapted for comic strip with panels. These refinements are simple ideas but they might even more improve the precision of the skew estimation.

3. RESULTS

During the DISEC13 competition [1], our method was evaluated and compared among twelve methods of ten different research teams. This evaluation was performed on 175 various/typical and representative scanned documents from various sources. These documents were randomly rotated in ten different orientations. 1550 document images among them were used for benchmarking dataset.

The performance evaluation is based on three different criteria. The first criterion is the Average Error Deviation (AED) which is the mean estimation error on the complete dataset. The second criterion is the Average Error Deviation computed only on the 80% best estimation (TOP80). This measure avoids algorithm to be penalized by outlayers in estimation (particular cases not well handled by the algorithm). The last criterion is the percent value of correct estimations (CE) (*i.e.* estimation where the error is lower than $0.1°$).

The results of the challenge, given in the competition report [1], show that:

- firstly our algorithm is very accurate with more than 77% of correct estimations,
- secondly our algorithm outperforms all other methods as our method is ranked first according to all evaluated criteria.

The following table shows the results for the 3 first methods. Complete results are provided in [1].

Method	AED (°)		TOP80 (°)		CE
	mean	std	TOP80	AED-TOP80	(%)
LRDE-EPITA-b	0.097	0.032	0.053	0.044	68.32
Ajou-SNU	0.085	0.10	0.051	0.034	71.23
Our	**0.072**	**0.06**	**0.046**	**0.026**	**77.48**

The percent value of correct estimations proves that the algorithm is very accurate. Furthermore the standard deviation is rather low, which proves the stability of the algorithm. Among all participating methods, it is the one whose score varies less between AED and TOP80 (the difference is only 0.026). This proves the robustness of the approach.

The method has proven to be very efficient and very competitive among many up-to-date methods. The only drawback is the resource consumption. The program is a research prototype and must be improved concerning both memory and time consumption. Especially, the computation of the KNN is in $O(n^2)$ according to the number of regions in the document. The next step for us will be to improve this resource consumption without loosing precision.

4. CONCLUSION

We have presented a new simple and accurate method for skew estimation in document images. Our method has two simple steps: a preprocessing using a KNN clustering and the skew estimation itself in the frequency domain.

Our method is robust and was ranked first at the DISEC'13 contest among many up to date methods. This proves the benefits of our algorithm.

Our main goal now is to improve the speed of the algorithm. We only have a research prototype of the algorithm, implementation of KNN must now be improved in order to run in real time and to be usable in a real complete large scale acquisition process which need very low executing time. However we also provide hints thanks to which we try to increase the precision of the estimation.

5. REFERENCES

[1] A. Papandreou, B. Gatos, G. Louloudis, and N. Stamatopoulos, "Icdar 2013 document image skew estimation contest (disec 2013)," *12th International Conference on Document Analysis and Recognition*, vol. 0, pp. 1444–1448, 2013.

[2] Jonathan J. Hull, "Document image skew detection: Survey and annotated bibliography," in *Document Analysis Systems II.* 1998, pp. 40 – 64, World Scientific.

[3] Sepideh Barekat Rezaei, Abdolhossein Sarrafzadeh, and Jamshid Shanbehzadeh, "Skew detection of scanned document images," *Proceedings of The International MultiConference of Engineers and Computer Scientists*, pp. 451 – 456, 2013.

[4] Nandini N., Srikanta Murthy K., and G. Hemantha Kumar, "Estimation of skew angle in binary document images using hough transform," *World Academy of Science, Engineering and Technology*, vol. 2, no. 6, pp. 44 – 49, 2008.

[5] Bin Yu and Anil K. Jain, "A robust and fast skew detection algorithm for generic documents.," *Pattern Recognition*, vol. 29, no. 10, pp. 1599 – 1629, 1996.

[6] Dan S. Bloomberg, Gary E. Kopec, and Lakshmi Dasari, "Measuring document image skew and orientation," in *Proc. SPIE 2422, Document Recognition II*, 1995, pp. 302 – 316.

[7] A. Papandreou and Basilios Gatos, "A novel skew detection technique based on vertical projections," in *International Conference on Document Analysis and Recognition*, 2011, pp. 384 – 388.

[8] Yue Lu and Chew Lim Tan, "Improved nearest neighbor based approach to accurate document skew estimation," in *International Conference on Document Analysis and Recognition*, 2003, pp. 503 – 507.

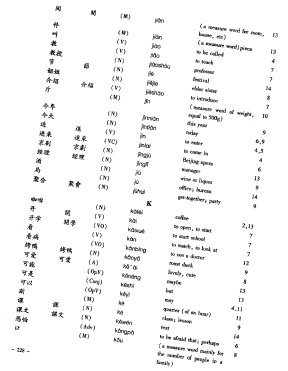
[9] Okun O., Pietikäinen M, and Sauvola J., "Robust document skew detection based on line extraction.," in *Proc. 11th Scandinavian Conference on Image Analysis*, 1999, pp. 457 – 464.

[10] L. Najman, "Using mathematical morphology for document skew estimation," in *SPIE Document Recognition and Retrieval IX*, 2004, pp. 182 – 191.

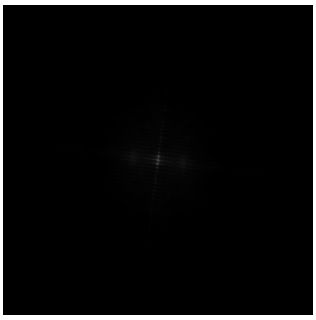
[11] G. S. Peake and T. N. Tan, "A general algorithm for document skew angle estimation.," in *IEEE International Conference on Image Processing*, 1997, pp. 230 – 233.

[12] Scott Lowther, Vinod Chandran, and Subramanian Sridharan, "An accurate method for skew determination in document images," in *Proceedings of the 6th Digital Image Computing: Techniques and Applications Conference*, Melbourne,Victoria, 2002, pp. 1–5, Australian Pattern Recognition Society.

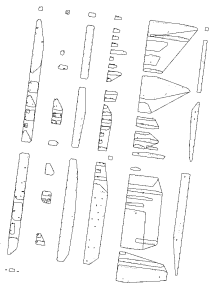
[13] Mandip Kaur and Simpel Jindal, "An integrated skew detection and correction using fast fourier transform and dct," in *International Journal of Scientific & Technology Res*, December 2013, vol. 2.



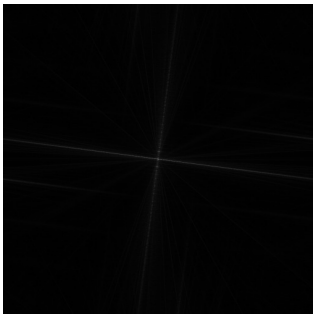
(a) Original document image



(b) Fourier transform of the original image

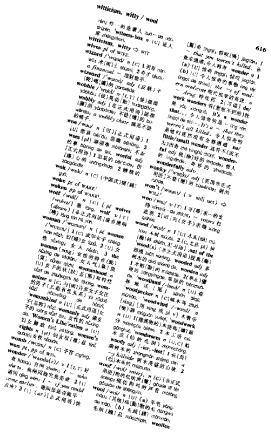


(c) Outline of convex hulls of all clusters made by KNN



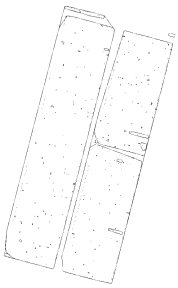
(d) Result of the Fourier transform

Fig. 1. Example of preprocessing step

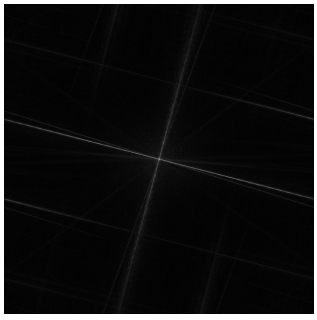


(a) Original document image

(b) Fourier transform of the original image



(c) Outline of convex hull of all clusters made by KNN



(d) Result of the Fourier transform

Fig. 2. Example of preprocessing step