Multi-oriented Text Extraction from Information Graphics

Falk Böschen
Kiel University
Olshausenstraße 40
24118 Kiel, Germany
fboe@informatik.uni-kiel.de

Ansgar Scherp
Leibniz Information Centre for Economics - ZBW
Düsternbrooker Weg 120
24105 Kiel, Germany
asc@informatik.uni-kiel.de

ABSTRACT

Existing research on analyzing information graphics assume to have a perfect text detection and extraction available. However, text extraction from information graphics is far from solved. To fill this gap, we propose a novel processing pipeline for multi-oriented text extraction from infographics. The pipeline applies a combination of data mining and computer vision techniques to identify text elements, cluster them into text lines, compute their orientation, and uses a state-of-the-art open source OCR engine to perform the text recognition. We evaluate our method on 121 infographics extracted from an open access corpus of scientific publications. The results show that our approach is effective and significantly outperforms a state-of-the-art baseline.

Categories and Subject Descriptors

I.7.5 [**Document and Text Processing**]: Document Capture—*Optical Character Recognition*

General Terms

Experimentation; Measurement

Keywords

infographics; OCR; multi-oriented text extraction

1. INTRODUCTION

Scientific publications often include information graphics (short: *infographics*) to visualize statistics, survey data, and research results in an easy to perceive way. Retrieval over infographics is typically based on the surrounding text. However, information graphics often contain textual information that is *frequently not present in surrounding text* [2]. Therefore, ignoring the textual information encoded in graphics discards a big opportunity to improve retrieval and understanding of the content.

Existing research on analyzing infographics such as [17, 7, 15, 5, 8, 4] just simply assume that a perfect OCR is readily available. However, as illustrated by the example in Figure 1, OCR on infographics poses several challenges that are insufficiently addressed by current solutions: (i) First, the text contained in infographics is often rotated at different angles to fit axes or graphical components. (ii) Second, the textual elements often have different fonts, sizes, and typographic emphases. (iii) Third, we find text elements that partially cover some graphical components in the infographics or have different background colors. This makes it difficult to identify the text elements and separate them from the graphical components.

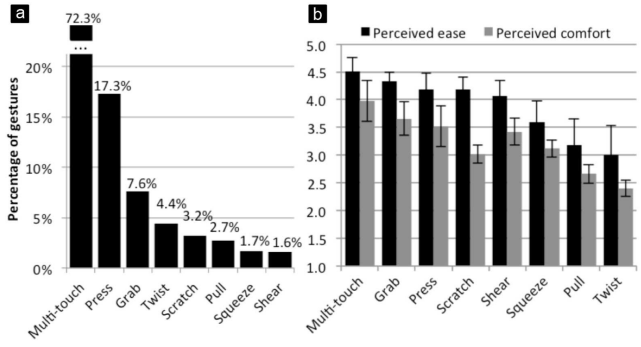

Figure 1: Challenges for text extraction from infographics (example taken from [18]).

In this paper, we propose an initial version of a novel infographics processing pipeline and conduct a first evaluation to proof its concept. The pipeline makes use of a combination of methods from data mining and computer vision to identify and extract text from infographics. The evaluation over 121 infographics extracted from an open access corpus of scientific publications demonstrates its effectiveness. It significantly outperforms a reasonable baseline applying the state-of-the-art OCR engine Tesseract[1].

The remainder of the paper is organized as follows: Subsequently, we discuss the related work. Section 3 presents our pipeline for text extraction. Section 4 specifies the experiment set-up and dataset used. The results are presented in Section 5 and discussed in Section 6, before we conclude.

2. RELATED WORK

Research on infographics typically focused on the classification task [17], i. e., determining the diagram type of a

[1] `https://code.google.com/p/tesseract-ocr/`

DocEng'15, September 8–11, 2015, Lausanne, Switzerland.

ACM 978-1-4503-3307-8/15/09 ...$15.00.
DOI: http://dx.doi.org/10.1145/2682571.2797092 .

graphic, or finding areas of text in the infographics [1, 14]. However, due to the huge variety of infographics, it shows to be hard to train such a classifier. Kataria et al. [7] address this issue by focusing on 2D plots and using the coordinates of the image regions as support to distinguish between text and graphics. Only few works deal with the identification of text elements in infographics with the purpose to extract the structural layout of text and graphics [9]. Chester [4] developed the VEM framework for extracting graphical symbols and their associated text but it is still very limited and only applicable to bar charts, pie charts, and line charts. The most advanced analysis techniques for infographics are provided by ReVision [15] and VIEW [5] that aim at reengineering data from bar charts and pie charts with the goal to render it in some other form. The SIGHT system extracts the core message of bar charts to make them accessible to visually impaired users [3].

All these approaches assume to have perfect OCR results from infographics. However, having high-quality OCR results from infographics is unrealistic. Thus, they use data sets with manually entered text labels. The lack of perfect OCR for infographics is also shown by the limitations of existing works on rotation-invariant OCR that require a single character in high resolution and perfectly cropped boundary [12] or are tailored to a particular problem, e.g. cartographic material [10].

Our approach aims to fill the gap by defining a robust text extraction pipeline for infographics containing text of various emphases, size, color, and multiple orientations.

3. TX PROCESSING PIPELINE

The pipeline for Text eXtraction from infographics (short: TX) comprises six steps as illustrated in Figure 2.

(1) Adaptive binarization and labeling. The first step performs a novel adaptive binarization based on a quadtree that hierarchically divides the infographic into tiles. For each tile, we determine an optimal binarization threshold by applying Otsu's method [11], which outperforms standard approaches based on a fixed threshold or histogram. For each tile, we apply the popular Sobel operator to determine the edges. We compute the Hausdorff distance over the edges of the current tiles and their parent tile. We further subdivide a tile if a certain empirical threshold is not reached. The final threshold is computed by averaging over all tiles. The resulting binary image is then labeled using the Connected Component Labeling method [13]. The output of this step is a set of labeled regions of the infographic.

(2) Grouping regions to text elements. The regions produced in step (1) are categorized into "text elements" and "graphic symbols". To this end, we calculate for each region a feature vector with the center of mass coordinates (based on the first order moments), bounding box (width and height), and mass-to-area-ratio. We apply the density-based clustering algorithm DBSCAN to categorize regions into "text elements" and noise ("graphic symbols" and others), since we do not know the number of clusters beforehand. Output of this step is a clustering where each cluster is a set of regions representing a candidate text element.

(3) Computing of text lines. Clusters created by DBSCAN do not necessarily represent text lines. Thus, we apply a second clustering based on a Minimum Spanning Tree (MST) on top of the DBSCAN results. The rationale is that regions belonging to the same text lines a) tend to be closer together (than other regions) and b) the edges between those regions are of similar orientation. For each cluster, the MST is build using the regions' center of mass coordinates. We compute a histogram over the angles between the edges in the tree and discard those edges that differ from the main orientation.

(4) Estimating the orientation of text lines. While the MST applied in step (3) can well produce potential text lines, it is not well suited for estimating the orientation of the text lines as it is computed on the center of mass coordinates. Thus, in the fourth step we apply a standard Hough line transformation to estimate the actual text orientation. This estimation is known to be error tolerant with regards to deviations in the orientations for small number of regions.

(5) Rotate regions and apply OCR. Based on the text lines computed in step (3) and their orientation in step (4), we crop rotated sub-images from the input infographic and apply OCR. In our current implementation, we use the state of the art OCR engine Tesseract without layout analysis (i. e., in single text block mode).

The last step (6) is the evaluation of the results, which is described in detail below.

4. EVALUATION SETUP

We compare the performance of our infographics analysis pipeline TX with a baseline based on Tesseract. We compute the results over 1-, 2- and 3-grams as well as words. A word is a series of alpha-numeric characters, i. e., it does not contain any white space characters. The n-grams are computed over the words. Below, we first describe the evaluation procedure and dataset used. Subsequently, we introduce our baseline and evaluation metrics.

4.1 Evaluation Procedure

We match the results of TX and the baseline with some gold standard to conduct our evaluation. The matching considers both the position of the text elements as well as their orientation in the infographics. We take each word from TX and baseline and calculate the intersection of its bounding box with the bounding box of all words in the gold standard. The pair with the maximum intersection is taken for evaluation. Therefore, we assign each extracted word to either one or zero words from the gold standard. All words from TX/baseline that were not assigned are considered to be false positives. All words from the gold standard without pairing are considered as false negatives. The n-grams are computed over the words of each set.

4.2 Dataset and Gold Standard

In our initial evaluation, we use a dataset of 121 infographics. We obtained them from a large corpus of 288,000 open access publications in the domain of economics we collected earlier. From this corpus, we extracted 200,000 candidate images by applying aggressive thresholds: We require a minimum width/height of 500 pixels as OCR is infeasible on smaller images. We also removed images of size above 2000 pixels as these were not individual graphics but rather scans of entire pages. From the candidate set, we randomly picked images - one at a time - and presented them to a human viewer to confirm that it is an infographic. For creating the gold standard, we have developed a web-based selection tool and manually annotated the infographics' text elements. Each text element contains the information about

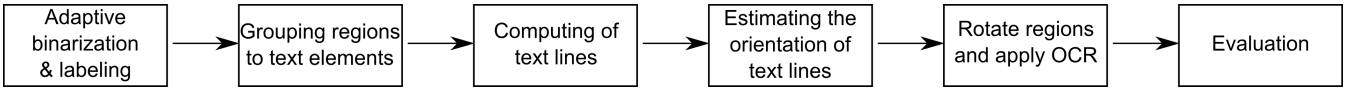

Figure 2: Novel processing pipeline for text extraction from infographics

its position, dimension, rotation, and alpha-numeric characters.

4.3 Baseline

Currently, there are no special tools freely available that are capable of performing text extraction from infographics. Related works such as rotation-invariant OCR [12, 10] are too limited to be applicable in a general context of extracting text from infographics (for details see Section 2). Thus, we use the state-of-the-art OCR engine Tesseract in its default mode, i. e., including layout analysis over the entire infographic, as a reasonable baseline. It is appropriate to apply Tesseract as baseline as it is capable of multi-oriented text extraction. Tesseract supports a rotation margin of $\pm 15°$ [16]. In addition, it can detect text rotated at $\pm 90°$.

4.4 Evaluation Metrics over Multisets

As mentioned in Section 4, we evaluate our pipeline over n-grams and words. For the n-grams, we apply the standard evaluation metrics precision (P), recall (R), and F_1-measure (F) as defined below:

$$P = \frac{|Extr \cap Rel|}{|Extr|},\ R = \frac{|Extr \cap Rel|}{|Rel|},\ F = \frac{2 \cdot P \cdot R}{P + R}$$

Extr refers to the n-grams as they are computed from text elements that are extracted from an infographic by TX and the baseline, respectively. *Rel* refers to the relevant n-grams from the gold standard. Both *Extr* and *Rel* are multisets, so we need to adjust the definitions of P and R. Multisets can appear insofar as the same n-gram can appear multiple times in both the extractions result from TX and the baseline as well as gold standard. To properly account for the number of occurrences of an n-gram in *Extr* or *Rel*, we define the counter function $\mathbf{C}_A(x) := |\{x|x \in A\}|$ (as an extension of a set indicator function) over a multiset A. For an intersection of multisets A and B, the counter function is defined as follows:

$$\mathbf{C}_{A \cap B}(x) := \min\{\mathbf{C}_A(x), \mathbf{C}_B(x)\} \quad (1)$$

Based on $\mathbf{C}_{A \cap B}(x)$, we define P and R for multisets:

$$P = \frac{\sum_{x \in Extr \cup Rel} \mathbf{C}_{Extr \cap Rel}(x)}{\sum_{x \in Extr} \mathbf{C}_{Extr}(x)} \quad (2)$$

$$R = \frac{\sum_{x \in Extr \cup Rel} \mathbf{C}_{Extr \cap Rel}(x)}{\sum_{x \in Rel} \mathbf{C}_{Rel}(x)} \quad (3)$$

In addition, it can happen that one of the sets *Extr* and *Rel* are empty. This refers to the situation when a) our TX pipeline or baseline do not extract a text where they should, i. e., $Extr = \emptyset$ and $Rel \neq \emptyset$. In this case, we define following Groot et al. [6] that $P := 0$ and $R := 0$ (false negative). In the situation b) where TX or the baseline find some text where they should not, i. e., $Extr \neq \emptyset$ and $Rel = \emptyset$, we define $P := 0$ and $R := 1$ (false positive).

For evaluating the results on level of individual words (i. e. sequences of alpha-numeric characters separated by blank or carriage return), we use standard Levenshtein distance.

5. RESULTS

Table 1 shows the average number of 1-, 2-, and 3-grams obtained from our extraction pipeline (TX), baseline (BL), and gold standard (GS) from the 121 infographics in our dataset. In addition, we show the average number of words extracted from the dataset and the average length of the words in number of characters. Standard deviations are provided in brackets.

	1-grams	2-grams	3-grams	Words	Length
TX	177.20 (128.20)	127.34 (100.51)	89.34 (79.35)	50.07 (31.95)	3.63 (2.69)
BL	106.30 (87.71)	80.17 (69.12)	60.79 (54.54)	25.21 (22.12)	4.15 (2.25)
GS	150.65 (122.28)	115.93 (103.09)	84.95 (85.61)	35.46 (22.24)	4.22 (1.48)

Table 1: Average number of n-grams and words extracted from the 121 infographics and average word length.

As one can see, our novel multi-oriented text extraction pipeline TX detects about 50% more n-grams and twice as many words as the baseline. However, the average length of the words is quite similar. In addition, TX extracts slightly more n-grams and words than the gold standard actually contains. Also, both the TX pipeline and baseline extract words that are shorter than the gold standard. Overall, we observe quite high standard deviations including the gold standard. Thus, the textual content of the 121 infographics is quite diverse.

The results of our comparison between the TX pipeline and baseline in terms of average P, R, and F measures (standard deviation in brackets) for the 1-, 2-, and 3-grams are reported in Table 2. In addition, we show the relative improvement of TX over the baseline, which is on average about 20% higher (all statistically significant except F-measure on 3-grams, details omitted here for reasons of brevity). In particular for R, the novel TX pipeline can achieve results of at least 35% above the baseline. One can also observe that in general the performance degrades from 1-grams to 3-grams. Finally, the results show a high standard deviation as well.

Regarding the Levenshtein distance, the results for our TX pipeline are on average 2.23 (SD=1.29). Thus, about two characters need to be changed for an exact match. For the baseline we report an average Levenshtein distance of 2.53 (SD=1.59). The difference of 0.3 is significant ($t(120) = 2.1, p < .04$) using a standard significance level of $\alpha = 5\%$.

Finally, for our pipeline one observes on average 12.94 false negatives (SD=17.88) as well as 49.87 false positives (SD=31.52). In comparison, we report on average 17.01 false negatives (SD=17.40) and 5.67 false positives (SD=9.42) for the baseline. Thus, the TX pipeline produces significant less

	n-gram	P	R	F
TX	1	.50 (0.41)	.68 (0.36)	.47 (0.39)
	2	.41 (0.41)	.60 (0.41)	.39 (0.39)
	3	.29 (0.38)	.49 (0.43)	.27 (0.36)
BL	1	.37 (0.36)	.48 (0.36)	.36 (0.35)
	2	.32 (0.35)	.44 (0.37)	.32 (0.35)
	3	.24 (0.32)	.36 (0.37)	.24 (0.32)
Diff.	1	35.32%	42.15%	28.95%
	2	28.15%	37.97%	20.86%
	3	18.00%	36.81%	11.94%

Table 2: Average P, R, F measures for TX and baseline.

false negatives ($V(120) = 4503.5, p < .01$). However, TX produces significantly more false positives than the baseline ($t(120) = -16.6, p < .001$).

6. DISCUSSION AND CONCLUSION

The results of our evaluation show the general effectiveness of the novel TX pipeline for the multi-oriented text extraction from infographics. This is especially documented by the increase in recall for the n-grams compared to the baseline. The increase in recall results from finding more text elements at different orientations. Also the precision increases, which results in an higher performance of TX as documented in the F-scores. We observe a quite high standard deviation in the results of both TX and baseline. This can be explained by the already high standard deviation in the gold standard. With other word, the infographics in our dataset are quite diverse in terms of the number of text elements they contain. Thus, the observed standard deviation is not an issue of our extraction pipeline (or the baseline) but an artefact of the dataset. One potential negative influence on the results of our evaluation is the decreasing number of 3-grams in the infographics. On average, we find only about half as many 3-grams than 1-grams. However, as the results in Table 1 show, there are still on average 80 3-grams in the gold standard. Enough to produce reasonable results.

Comparing the extracted words using Levenshtein distance shows that we can detect them significantly better than the baseline. Our TX engine has less false negatives, i. e., it extracts more text elements from the gold standard than the baseline can do. However, the TX pipeline makes more mistakes in terms of extracting text elements where there are none in the gold standard. This is documented in Table 1, where one can see that our TX pipeline extracts on average more text elements as there are actually in the infographics (defined by the gold standard). A detailed analysis of these false positives shows that those falsely extracted text elements mostly contain special characters. Thus, it should be possible to remove these false positives in a future extension of our work.

So far, we have used the state-of-the-art OCR software Tesseract in our pipeline. In the future, we will also apply alternative OCR engines like Ocropus[2]. Most importantly, we will extend the dataset used in this paper by annotating infographics at large scale using a crowd-sourcing approach.

Acknowledgment. We thank Chifumi Nishioka for collecting the open access publications used in our experiments.

[2] `https://github.com/tmbdev/ocropy`

7. REFERENCES

[1] P. Agrawal and R. Varma. Text extraction from images. *IJCSET*, 2(4):1083–1087, 2012.

[2] S. Carberry, S. Elzer, and S. Demir. Information graphics: an untapped resource for digital libraries. In *SIGIR*, pages 581–588. ACM, 2006.

[3] S. Carberry, S. E. Schwartz, K. F. McCoy, S. Demir, P. Wu, C. Greenbacker, D. Chester, E. Schwartz, D. Oliver, and P. Moraes. Access to Multimodal Articles for Individuals with Sight Impairments. *TiiS*, 2(4):21:1–21:49, January 2013.

[4] D. Chester and S. Elzer. Getting Computers to See Information Graphics So User Do Not Have to. In *Foundations of Intelligent Systems*. Springer, 2005.

[5] J. Gao, Y. Zhou, and K. E. Barner. VIEW: Visual information extraction widget for improving chart images accessibility. In *Image Processing*. IEEE, 2012.

[6] P. Groot, F. van Harmelen, and A. ten Teije. Torture tests: A quantitative analysis for the robustness of knowledge-based systems. In *EKAW*, 2000.

[7] S. Kataria, W. Browuer, P. Mitra, and C. L. Giles. Automatic extraction of data points and text blocks from 2-dimensional plots in digital documents. In *Advancement of Artificial Intelligence*. AAAI, 2008.

[8] Z. Li, S. Carberry, H. Fang, K. McCoy, and K. Peterson. Infographics Retrieval: A New Methodology. In *Natural Language Processing and Information Systems*. Springer, 2014.

[9] Z. Li, M. Stagitis, S. Carberry, and K. F. McCoy. Towards retrieving relevant information graphics. In *SIGIR*. ACM, 2013.

[10] R. Mariani, M. P. Deseilligny, J. Labiche, and R. Mullot. Algorithms for the hydrographic network names association on geographic maps. In *Document Analysis and Recogn.* IEEE, 1997.

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

[12] P. M. Patil and T. R. Sontakke. Rotation, scale and translation invariant handwritten devanagari numeral character recognition using general fuzzy neural network. *Pattern Recogn.*, 40(7):2110–2117, July 2007.

[13] H. Samet and M. Tamminen. Efficient component labeling of images of arbitrary dimension represented by linear bintrees. *IEEE TPAMI*, 10(4):579–586, 1988.

[14] J. Sas and A. Zolnierek. Three-stage method of text region extraction from diagram raster images. In *CORES*. Springer, 2013.

[15] M. Savva, N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, and J. Heer. ReVision: Automated Classification, Analysis and Redesign of Chart Images. In *UIST*, pages 393–402. ACM, 2011.

[16] R. Smith. A simple and efficient skew detection algorithm via text row accumulation. In *Document Analysis and Recogn.*, volume 2, Aug 1995.

[17] F. Wang and M.-Y. Kan. NPIC: Hierarchical synthetic image classification using image search and generic features. In *Image and Video Retrieval*. Springer, 2006.

[18] M. Weigel, V. Mehta, and J. Steimle. More than touch: understanding how people use skin as an input surface for mobile computing. In *CHI Conference*, pages 179–188. ACM, 2014.