# Geometric Layout Analysis in a Wearable Reading Device for the Blind and Visually Impaired

Roman Guilbourd and Raul Rojas

Free University Berlin
{guilbour,rojas}@mi.fu-berlin.de

**Abstract.** Blind and visually impaired people can use a mobile device for accessing printed information, which is ubiquitous in everyday life. Thus, there is a need for a mobile easy-to-use reading device, capable of dealing with the complexity of the outdoor environment. In this paper a wearable camera based solution is presented, aiming at improving the performance of existing systems through the use of an integrated approach for the document processing. This particular publication covers the segmentation phase of the processing chain as well as geometric analysis of the layout. Using a highly efficient approach we were able to overcome the limitations of a mobile computing environment without compromising on the robustness of the result. In order to demonstrate the advantages of the presented algorithm for the specific field of application we compare its output to the results obtained by a state-of-the art commercial solution.

**Keywords:** Wearable device, healthcare assistance, OCR, document processing.

## 1 Introduction

In our project we are developing an eyewear system for blind and visually impaired people, which would automatically detect, localize and process text documents in both in- and outdoor environments. Most of the currently available standard solutions in this area require a significant amount of user involvement, in particular during the acquisition phase, when the capturing device has to be oriented towards the document. This might be one of the major reasons for the continuing prevalence of stationary reading devices that was indicated by our survey.

Apart from the stationary solutions there exist numerous smartphone-based aids, which, however, can be difficult to handle for people with limited spatial vision due to its handheld nature. The mobility requirement poses a considerable challenge in terms of robustness and efficiency of the algorithms utilized in the document processing chain. Table 1 illustrates the additional effort necessary to achieve a result that would be comparable with that of a scanner-based reading device. Compounding the problem are the limited computational resources of a mobile platform, which are a result of the miniaturization effort. Therefore, an efficient and highly integrated approach was developed through optimization of the entire processing chain as a whole. Several steps of the procedure have already been presented in [1]. This current paper is focused on the segmentation phase showing how different parts of the chain interact in order to

G. Memmi and U. Blanke (Eds.): MobiCASE 2013, LNICST 130, pp. 90–97, 2014.

reduce overall processing time. In the following we will briefly introduce the proposed document processing routine and then discuss image segmentation and text detection in more detail, demonstrating the links between the different stages of the pipeline. In conclusion, some evaluation results are presented and the benefits of the proposed algorithm for the particular application are explained.

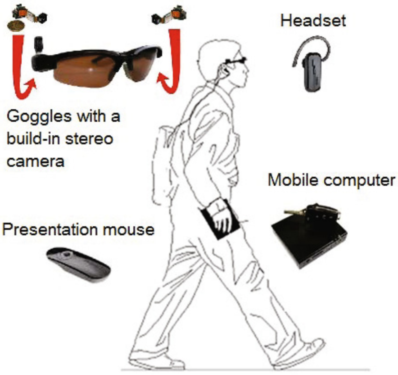

**Fig. 1.** System components

**Table 1.** Stationary vs. Mobile Systems

| Stationary | Mobile |
|---|---|
| Position of the document is determined | Text localization required |
| Lighting conditions are controllable(scanner) | Calibration required |
| Distance to the document is fixed | Auto-focusing required |
| Distortion is negligible | De-warping required |
| No motion artifacts | Prevention required |

## 2 Related Work

Several multiresolution approaches for segmentation of document images have been suggested by different authors. In [2] the performance of wavelet packet spaces is analyzed in terms of sensitivity and selectivity for the classification of textures in general. In particular, a comparison of energy and entropy based signatures is performed and, as a result of an experimental analysis, energy representations computed from the standard set of wavelet nodes is shown to be sufficient for texture classification. In the work of Etemad at al. [3] wavelet based features are utilized for the specific case of textures presented in a printed document with four kinds of layout elements considered as different classes: text, image, graphics and blank areas. In the suggested method every sub-block of a document image is pre-classified using a multilevel feed-fordward neural network. After that the results of the classification are combined by means of a soft decision integration approach. The author also mentions, without going into more detail, that wavelet based features can indicate font size of characters contained in a text block. Distribution of the wavelet coefficients in high frequency bands are analyzed in [4] in order to identify text and non-text regions of a document. The presented classification approach follows the assumption that a histogram of the coefficients from a text area should be of a highly discrete nature. Several authors [5] [6] have proposed an approach, where the quadtree-representation of an image in the position domain is combined with the FWT-decomposition of the signal in the frequency domain. The problem of font size estimation basing on the distribution of the FTW-coefficients is addressed in [7]. The author introduces a method for detection of text areas using third andfourth central moments of the corresponding scalogram pdf. However, no information is provided on the efficiency of the implementation.

## 3 System Description

The system presented in this paper comprises a goggle with a build-in stereo camera, a mini ITX computer featuring a 1.6 GHz CPU, a wireless presentation mouse and a headset (as seen in figure 1). By using a stereo camera we were able to increase the field of view of the device, thus alleviating the orientation difficulties, while at the same time tackling some of the most prominent problems in the field of document processing like text tracking, de-warping of text regions and exact focusing [1]. The whole chain can be divided into two stages: capturing phase and processing phase. The capturing phase serves to detect and capture textual information while ensuring a suitable quality of the image as required for successful character recognition. In its initial state the system scans the environment for text objects providing the user audible feedback upon detection. After the notification the user can either initiate the processing of the detected textual information or go back to the initial loop. This basic functionality can be accomplished by using just two buttons on the presentation mouse, whereas two more buttons are needed to access some advanced services such as permanent storage of the recognized text and text navigation. Pending a decision, the systems keeps tracking the detected regions and produces a warning, if a text block has reached a boundary of the visible image area. After the user has initiated the processing, optimal camera settings and orientation are determined utilizing the measurements from the tracking phase.

The second phase includes a preprocessing of the stereo image, an actual OCR (optical character recognition), a TTS (text-to-speech) transformation and the output of the result. It starts with a segmentation of the two high resolution images as well as a classification of the separated regions. Unlike in the case of the real-time text detection algorithm from the capturing phase, the completeness and accuracy of the result are of crucial importance for the layout analysis. In order to meet the requirements a novel document image segmentation algorithm based on the Fast Wavelet Transformation (FWT)[8] and quadtrees [9] was developed and integrated into the processing chain. After the segmentation is completed, a binarization of the regions [10] and extraction of the text lines is performed in a combined approach involving connected component analysis. The set of the extracted components is then analyzed in order to reliably identify text regions, which are to be further processed. First, the structure of the documents in the images is recognized and every text block is labeled as a component of the logical layout. Basing on the result of the layout analysis, a proper reading order for the text regions is obtained and an OCR is accomplished according to that order. Using the text specific features extracted in the previous steps, de-warping and stitching of the regions can be performed [1] prior to the character recognition, when necessary. Finally, an OCR and a TTS-conversion are carried out using commercial solutions [11]. The output is initiated immediately after the first text block is ready, while the processing is continued in the background.

# 4 Image Segmentation and Text Detection

## 4.1 Specific Requirements

Segmentation of the images is the first step of the presented document processing chain, so no information about the content of the images is available at the time. While a reliable recognition of text regions depends upon a correct segmentation, some data on text specific properties are required to support the segmentation. In order to be able to deal with this mutual dependency a few general assumptions about the text regions are made prior to the segmentation step:

- (initially)bimodal distribution of the pixel values in document areas
- (nearly) uniform distribution of character heights in a single text region
- (nearly) uniform distribution of inter-character and interline spacing in a single text region
- the distance between two lines in a text block is relative to the average inter-character spacing in the lines and is at least twice as large as its maximum
- the distance between two text regions is relative to the average inter-line spacing and is at least twice as large as its maximum

A coarse estimation of the possible font size presented in the segments is required, since some of the subsequent steps of the processing such as text line extraction and binarization [10] rely on it.

The main motivation of developing a mobile reading device is that it could be used in an outdoor environment, where the user is confronted with a wide variety of different presentations of textual information. Compared to the special case of printed documents, distinguishing between text and non-text elements, that are embedded in a dynamic natural environment, can be a much more difficult. Unlike most of the related works presented above, our segmentation algorithm does not include a comprehensive classification of the segments. After connected components of a region are extracted as part of the binarization procedure, identification of text areas can be accomplished in a more straightforward way.

## 4.2 Image Segmentation

For reasons of robustness, reliability and efficiency a multiresolution FWT-based approach is utilized for image segmentation. In the position domain a quadtree decomposition is performed [12], so each stage of the FWT has a corresponding level in the tree. The pattern of distribution of the energies over the scales

$$Energy_i = \sum_j \left|\alpha_i^j\right|^2,$$

with $\alpha^i$ denoting the FWT coefficients from $i$th decomposition level, is analyzed to determine the similarity between subregions. Energy representations computed from the standard set of wavelet nodes have been shown to be sufficient for texture classification [2]. In addition, locations of the peaks in the energy distribution indicate such important text specific properties of regions as possible font size (as seen in figure 2) and text flow direction, that are both required for the subsequent steps of the processing. The feasibility of the conclusions follows directly from the scaling property as well as energy conservation property of the FWT [8], since differences in the font size can be considered as shifts in the frequency spectrum of the image signal.

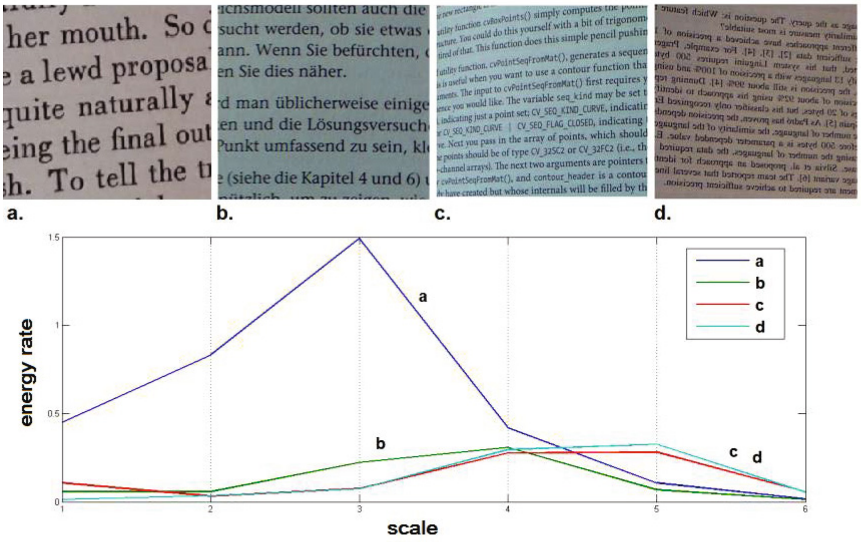

**Fig. 2.** Correspondence between font size and peaks in the FWT energy distribution

The segmentation of the image is carried out as follows:

1. Energy distribution signature of the subregions is computed up to the corresponding decomposition level.
2. Homogeneous background areas are identified using a variance threshold.
3. Neighboring subregions are incrementally merged taking into account the background areas.

The recognition of the homogeneous subregions occurs by means of the universal threshold $\lambda_u$=$\sqrt{2lnN}\hat{\sigma}$, that was first introduced and frequently utilized for noise reduction [15]. Here is $N$ - number of the FTW-coefficients in the region and $\hat{\sigma}$ - median absolute deviation (MAD) of all coefficient values on the respective decomposition level. Since computation of MAD is quite a time consuming operation, variance $\sigma$ of the coefficients is used to approximate $\hat{\sigma}$ instead:

$$\sigma^2\left(\alpha_i\right) = E\left(\alpha_i^2\right) - \left(E\left(\alpha_i\right)\right)^2 = \frac{1}{N_i} Energy_i - \left(E\left(\alpha_i\right)\right)^2,$$

$E$ denotes here the expected value whereas $Energy$ denotes the overall energy of the coefficients $\alpha$ on the level $j$. This way, no additional time is needed to determine the threshold $\lambda_u$, however, the accuracy of the classification can be compromised by potential outliers. At this point of computation it is especially important to avoid false positive results (as seen in figure 3 ), so that the threshold is scaled by a factor $\gamma < 1$.

Nodes in the tree that correspond to the homogeneous subregions are then marked as background. According to the assumptions about text regions presented above, two text blocks should be separated by an interval at least twice the width of the inter-line spaces of the blocks. As a consequence, homogeneous areas from within the regions should be located on a lower tree level than blank areas from the intermediate spaces. Moreover, the assumptions also imply, that the width of a spacing interval should correspond to the font size in the surrounding text blocks and therefore to the "peak" scale of the energy distributions in the neighboring tree nodes. With that in mind we can merge detected background areas and filter out most of the false positives by applying following strategy:

1. Starting with the leaves follow the tree up until the "peak" scale level is reached.
2. If a "peak node" has any background neighbors mark them as border regions.
3. Propagate the signature $\{h\left(R_{max}\right), o\left(R_{max}\right)\}$ of the "peak nodes" down to the leaves, where $h$ denotes distribution of the energy over the scales, $o$ denotes distribution of the energy over the FWT subbands corresponding to the ridge orientation and all feature vectors are assumed to be normalized.
4. Traverse the tree one more time and mark all the descendants of the border region according to the scale of the "peak-node".
5. Apply marker-controlled watershed segmentation algorithm [13].
   - Use the mismatch between the propagated "peak node" scale and the current level of the marked border regions as the segmentation function.
   - Use the signature-based similarity function to obtain foreground markers: $sim\left(R_1, R_2\right) = h\left(R_1\right) \cdot h\left(R_2\right) + o\left(R_1\right) \cdot o\left(R_2\right)$.
   - Use the "peak-node"-scale to obtain background markers.
   - for each foreground marker perform a region growing operation on the corresponding tree level.

### 4.3 Text Detection

Utilizing the approach presented above we not only managed to perform a text specific segmentation of the image, but also to estimate the font size in the segments, thus being able to parameterize the scale sensitive binarization algorithm [10]. The subsequent analysis of the extracted regions is carried out on the binarized image of the positive translation energies. The blocks are then scanned in the direction perpendicular to the estimated orientation of the text lines (as determined by means of the FWT energy distribution) and foreground objects are extracted using [16]. Provided that the measurements of the characters are distributed uniformly, it can be easily shown, that the characters of a single line are detected in the sequential order as long as the orientation error does not exceed certain limits, e.g. 45° in case that the maximum height difference and the maximum gap size between two neighboring characters are both limited to the height of the smaller character. During this operation the contours of the components are marked to avoid rediscovering already known elements and the size of the elements is estimated by bounding volumes. After the scanning process is completed, the coefficient of variation of heights of the bounding boxes is computed to verify the "uniformity" condition: $\sigma\left(h_s\right)/E\left(h_s\right) < 2$, where $\sigma\left(h_s\right)$ - the standard deviation and $E\left(h_s\right)$ - the expected value. In case the requirement is not fulfilled, the segment is classified as non-text and discarded. Otherwise, the characters are arranged into lines and the distances between the characters are analyzed in a similar way to how character size distribution was analyzed before. Since the characters of a line are discovered in an order close to the sequential, only few characters from the current end of each line need to be considered as potential predecessors of a text line element to be sorted.

## 5 Result and Evaluation

In this paper a document processing chain for a wearable reading system is presented of which the segmentation stage is discussed in more detail. The overall result of the initial layout analysis includes positions, measurements and specific characteristics of the text blocks, text lines and single characters contained in the image. Thanks to the integrated approach, we were able to incorporate the binarization algorithm into the segmentation operation for their mutual benefit. The ability of the segmentation algorithm to systematically differentiate between inter-line and inter-region spacing basing on the font size is particularly important in outdoor environments, where text regions are tightly embedded into the scene and could be partly obscured by non-text objects. A special effort was made in order to optimize the procedure in terms of the efficiency. The execution time average for the combined approach including segmentation, binarization, text line extraction and classification of the segments given two 8 Mpx images on a 1.6 GHz processor is at 1 ms per extracted component. After these steps the amount of text to be recognized and output at a single round of reading can be dynamically determined with a precision up to a single word, depending on the estimated processing time for the subsequent region. The *response* time of the system for a document containing 10000 symbols is around 20 s, whereas the overall processing time can exceed two minutes.

OmniPage Capture SDK 16, which is used to perform character recognition in our system, also features geometric layout analysis making it possible to evaluate the

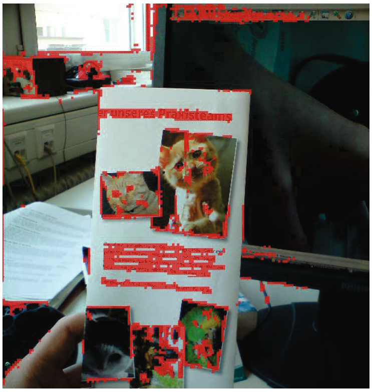

**Fig. 3.** Segmentation result

**Table 2.** Evaluation result of the algorithm

| | **Our approach** | **OmniPage** |
|---|---|---|
| **MediaTeam DB**<br>1793 regions | 91%<br>1625 | 95%<br>1702 |
| **Additional images**<br>128 regions | 77%<br>99 | 31%<br>40 |

accuracy of the proposed approach. Two different data sets were used for evaluation purposes: a publicly accessible document database [14] presenting a wide variety of layout/font types and 14 images of severely distorted documents made outdoors under realistic lighting and environmental conditions. The rating was explicitly designed to detect oversegmentation and region merging by evaluation of coordinates of the extracted bounding boxes. While in the public data set ca. 90% (as seen in Table 2) of the text regions were localized and classified correctly, the success rate for the second set containing document images of a poorer quality was about 80%. In contrast, the performance of the OmniPage layout recognition module on images of distorted documents dropped dramatically, even though the recognition rate on the images from the MediaTeam document database was slightly higher than that of our implementation.

## References

1. Guilbourd, R., Yogev, N., Rojas, R.: Stereo camera based wearable reading device. In: Proceedings of the 3rd Augmented Human International Conference, vol. 1. ACM (2012)
2. Laine, A., Fan, J.: Texture Classification by Wavelet Packet Signatures. IEEE Trans. Pattern Anal. Mach. Intell. 15, 1186–1191 (1993)
3. Etemad, K., Doermann, D.S., Chellappa, R.: Multiscale Segmentation of Unstructured Document Pages Using Soft Decision Integration. IEEE Trans. Pattern Anal. Mach. Intell. 19(1), 92–96 (1997)
4. Li, J., Gray, R.M.: Context-based multiscale classification of document images using wavelet coefficient distributions. IEEE Trans. Image Process. 9, 1604–1616 (2000)
5. Lee, S.-W., Ryu, D.-S.: Parameter-Free Geometric Document Layout Analysis. IEEE Trans. Pattern Anal. Mach. Intell. 23(11), 1240–1256 (2001)
6. Cheng, H., Bouman, C.A.: Multiscale bayesian segmentation using a trainable context model. IEEE Trans. Pattern Anal. Mach. Intell. 10(4), 511–525 (2001)
7. Gupta, P., Vohra, N., Chaudhury, S., Joshi, S.D.: Wavelet Based Page Segmentation. In: Indian Conf. on Computer Vision, Graphics and Image Processing, pp. 20–22 (2002)
8. Rioul, O., Vetterli, M.: Wavelets and Signal Processing. Signal Processing Magazine 8(4), 14–38 (1991)
9. Finkel, R., Bentley, J.L.: Quad Trees: A Data Structure for Retrieval on Composite Keys. Acta Informatica 4(1), 1 (1974)
10. Block, M., Rojas, R.: Local Contrast Segmentation to Binarize Images. In: International Conference on the Digital Society, vol. 1(1) (2009)
11. OmniPage Capture SDK 16, Nuance Communications, Inc.
12. Choi, H., Baraniuk, R.G.: Multiscale image segmentation using wavelet-domain hidden Markov models. IEEE Trans. Image Process. 1309–1321 (2001)
13. Najman, L., Schmitt, M.: Geodesic saliency of watershed contours and hierarchical segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 18(12), 1163–1173 (1996)
14. Sauvola, J., Kauniskangas, H.: MediaTeam Document Database II. CD-ROM collection of document images, University of Oulu, Finland, http://www.mediateam.oulu.fi/MTDB/index.htm
15. Donoho, D.L., Johnstone, I.M.: Ideal Spatial adaptation via wavelet shrinkage. Biometrika 81, 425–455 (1994)
16. Suzuki, S., Abe, K.: Topological structural analysis of digitized binary images by border following. Computer Vision, Graphics, and Image Processing 30(1), 32–46 (1985)