Computing and Informatics, Vol. 34, 2015, 383–401

STATISTICS ORIENTED PREPROCESSING OF DOCUMENT IMAGE

Darko Brodić, Čedomir A. Maluckov

University of Belgrade
Technical Faculty in Bor
Vojske Jugoslavije 12, 19210 Bor, Serbia
e-mail: dbrodic@tf.bor.ac.rs

Liangrui Peng

Tsinghua University
Department of Electronic Engineering
Beijing 100084, P.R. China
e-mail: penglr@tsinghua.edu.cn

Abstract. Old printed documents represent an important part of our cultural heritage. Their digitalization plays an important role in creating data and metadata. The paper proposed an algorithm for estimation of the global text skew. First, document image is binarized reducing the impact of noise and uneven illumination. The binary image is statistically analyzed and processed. Accordingly, redundant data have been excluded. Furthermore, the convex hulls are established encircling each text object. They are joined establishing connected components. Then, the connected components in complementary image are enlarged with morphological dilation. At the end, the biggest connected component is extracted. Its orientation is similar to the global orientation of text document which is calculated by the moments. Efficiency and correctness of the algorithm are verified by testing on a custom dataset.

Keywords: Document image analysis, image analysis, moments, optical character recognition, statistical analysis, text skew

Mathematics Subject Classification 2010: 68U10, 62H35, 65D18, 46N30

1 INTRODUCTION

Old printed documents like manuscripts, printed books, historical maps, newspapers and journals [1], and monographs represent an important part of our cultural heritage. It is quite common for such documents to suffer from slight degradation problems. They include variable illuminated background, ink seeping, smear and strains. These phenomena make the process of image preprocessing difficult producing recognition errors. Therefore, appropriate methods should be developed to remove such noise from historical documents during digitalization.

In document automatic recognition systems, the quality of the input image is crucial to the final performance. During the scanning process, the effects that include noise and skew are unavoidable. These components can damage the image and decrease the performance of the recognizer. Skew correction plays an important role in the image preprocessing. A small inclination in document image can interfere in the layout analysis and consequently, in the rest of the process. Hence, the appearance of the skew in document image leads to the significant failures in the optical character recognition (OCR) system. It is true, because the OCR system shows a considerable sensitivity to any skew appearance in document image. Hence, the skew detection is a key element in digital image analysis and processing [2, 3, 4].

The old printed documents are characterized with shape regularity as any other printed text [2]. Hence, it contains letters with similar sizes. The distance between text lines is adequate, which facilitates separation of text lines. The orientation of the text lines is similar. It constitutes the uniform text skew, which is called the global text skew. Due to the nature of the printed text, each object in a document image incorporates similar or the same text skew attribute. In order to exclude noise or redundant elements, a preprocessing step is mandatory. This type of preprocessing step is sometimes called filtering [5]. It includes decision making process of excluding data from document image based on geometrical filters. In order to discard components which are likely to correspond to non-textual objects, some attributes of text are specified and classified [5, 6]. The selection of filters is based on the analysis of some internal features, which are inherent to itemize connected component and its neighborhood. Its task is to eliminate the connected components which have a low probability of being text objects. Typically, these features are area, density, width to height ratio [5], or some other inherent geometrical attributes [6, 7].

Different techniques that incorporate morphological and geometrical transformations have proven to be very efficient methods for a text skew estimation. In such a way, morphological transformation means the use of image dilation and erosion along with structuring element characterized by different length [8, 9, 10]. It implicit methods that incorporate morphological transformation proved correctness with low absolute error values. Furthermore, such type of method sometimes incorporates moment analysis. Moments measure the pixel distribution in the image [11, 12, 13, 14]. Accordingly, they are sensitive to the rotation, which makes them suitable for text skew identification. However, this method is applicable only to the single skew estimation [6], i.e. global text skew. A method based on geometrical transformation is proposed in [16]. It detects a dominant skew in document images using piecewise covering of document objects by parallelograms. The method proved to be very accurate, but the skew angle range is limited to 15°. The improvement of the method is proposed in [17]. A much more complicated method is proposed in [6]. It includes a complex preprocessing stage, which is established by multi-step geometrical filtering as well as multi-stage decision making process. Furthermore, the global text skew is identified with cross-correlation method, which is applied to the remained connected components. At the end, the local text skew is calculated by the least square method. This technique performs local skew estimation, reliable text localization without restriction on a skew angle value. However, this is a multi-stage and computer expensive method.

In the paper, we propose a robust algorithm suitable for the recognition of text skew in the old printed documents mainly made by typewriter like letters, technical notes, etc. Such type of documents are typically characterized with dominant skew, i.e. global text skew. The first stage of the algorithm is binarization. It reduces the impact of noise and uneven illumination creating the binary document image. In order to exclude redundant data, binary image is statistically analyzed and processed. Accordingly, the considerable attention is paid to statistics oriented preprocessing of binary document image. In this step, text elements classified in the group that represents the smallest and largest text heights are excluded. The exclusion of items is based on the percentile level of the object heights distribution. Furthermore, the convex hulls are established encircling each text object. They are filled inside and later joined to create connected components. Then, the connected components in complementary image are enlarged with morphological dilation. At the end, the biggest connected component is extracted. Its orientation is similar to global orientation of text document, which is determined by the moments [5, 14]. Efficiency and correctness of the algorithm are verified by testing on custom dataset.

Organization of this paper is as follows. Section 2 describes the proposed algorithm. Section 3 defines text experiments. Section 4 shows test result and discusses it. Section 5 makes conclusions.

2 PROPOSED ALGORITHM

The proposed algorithm identifies global text skew of the printed documents, which represents the dominant text skew of the whole document. It consists of the steps that follow:

1. Binarization.
2. Statistical analysis and preprocessing.
3. Text skew estimation based on the extended convex hulls.

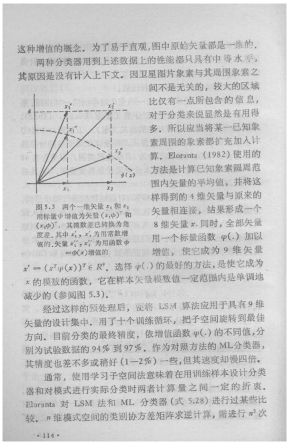
Binarization reduces the uneven illumination of the image background and creates binary document image. In order to exclude noise and redundant data, statistical analysis based on the object size distribution is performed. Accordingly, the lowest value of covariance gives the best results of the estimated text skew deviation. Hence, it shows the way how to choose the optimal percentile distribution ranges of text object heights in document image to discard redundant data. Then, the algorithm for text skew estimation based on convex hulls is applied to that image creating connected components. They are enlarged with a morphological operation. The biggest connected component is extracted. Its skew is detected with moments. At the end, the starting document image is rotated with the detected text skew.

2.1 Binarization

The binarization method adopts both global [18] and local adaptive thresholds [19]. It consists of three steps. First, multiple candidate thresholds are found via calculating the flatness of the histogram of the original gray image. Those pixels, which represent background or foreground pixels definitely, are selected by comparing with the largest threshold or the preset foreground threshold. The remained gray pixels will be binarized by the adaptive thresholding method in the next step. Second, we get the binarized results of the remained gray pixels by calculating adaptive threshold according to the mean and variance of the intensity levels in a local window. The statistical parameters such as average run-length are calculated for the binarized image. If the statistical parameters are normal values for text images, for example, the average run-length is in the range of normal average stroke widths, the binarized image is accepted. Otherwise, the original gray image will be binarized in the final step. Finally, a global threshold will be calculated by a histogram analysis. The multiple candidate thresholds found in the first step are checked one by one if the binarized image has normal value of the average run-length. If more than one threshold are feasible, the largest one is selected. If none of the candidate thresholds meets the requirement of the average run-length, the intensity level G which has the largest found value in the histogram, and G-E is taken as the global threshold. The value of E is determined empirically. In our experimental settings, E is assigned as 1. As a result, document image is transformed into a binary matrix $\mathbf{B}$ with M rows, N columns and two intensity levels $\{0, 1\}$. Figure 1 shows the document image before and after binarization process.

2.2 Statistical Analysis and Preprocessing

Each distinct element in the binary image represents the object commonly called a connected component CC. They usually correspond to separated characters, fragment of characters or even some noise elements. The primary task of the preprocessing stage is to disqualify superfluous data. In order to exclude redundant data from the image, the analysis of the object size distributions is performed. Figure 2 shows the histograms of object heights $\mathrm{CC}_{\mathrm{height}}$ distributions in the document image for different skew angles. The histograms are plotted with the class which width is equal



这种增值的概念。为了易于直观，图中原始矢量都是一维的。

两种分类器用到上述数据上的性能都只具有中等水平，其原因是没有计入上下文。因卫星图片象素与其周围象素之间不是无关的，较大的区域比仅有一点所包含的信息，对于分类来说显然是有用得多。所以应当将某一已知象素周围的象素都扩充加入计算。Eloranta (1982) 使用的方法是计算已知象素圆周范围内矢量的平均值，并将这样得到的 4 维矢量与原来的矢量相连接，结果形成一个 8 维矢量 x. 同时，全部矢量用一个标量函数 $\varphi(.)$ 加以增值，使它成为 9 维矢量 $x' = (x^T \varphi(x))^T \in R^9$。选择 $\varphi(\ )$ 的最好的方法，是使它成为 x 的模数的函数，它在样本矢量模数值一定范围内是单调地减少的（参阅图 5.3）。

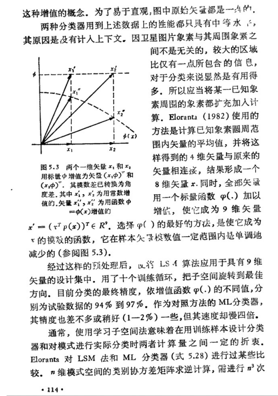
图 5.3 两个一维矢量 x_1 和 x_2 用标量 ϕ 增值为矢量 $(x_1\phi)^T$ 和 $(x_2\phi)^T$. 其模数差已转换为角度差. 其中 x_1', x_2' 为用常数增值的. 矢量 x_1'', x_2'' 为用函数 $\phi=\phi(x)$ 增值的

经过这样的预处理后，再将 LSM 算法应用于具有 9 维矢量的设计集中。用了十个训练循环，把子空间旋转到最佳方向。目前分类的最终精度，依增值函数 $\varphi(.)$ 的不同值，分别为试验数据的 94% 到 97%。作为对照方法的 ML分类器，其精度也差不多或稍好（1—2%）一些，但其速度却慢四倍。

通常，使用学习子空间法意味着在用训练样本设计分类器和对模式进行实际分类时两者计算量之间一定的折衷。Eloranta 对 LSM 法和 ML 分类器（式 5.28）进行过某些比较。n 维模式空间的类别协方差矩阵求逆计算，需进行 n^3 次

• 114 •

a)



这种增值的概念。为了易于直观，图中原始矢量都是一维的。

两种分类器用到上述数据上的性能都只具有中等水平，其原因是没有计入上下文。因卫星图片象素与其周围象素之间不是无关的，较大的区域比仅有一点所包含的信息，对于分类来说显然是有用得多。所以应当将某一已知象素周围的象素都扩充加入计算。Eloranta (1982) 使用的方法是计算已知象素圆周范围内矢量的平均值，并将这样得到的 4 维矢量与原来的矢量相连接，结果形成一个 8 维矢量 x. 同时，全部矢量用一个标量函数 $\varphi(.)$ 加以增值，使它成为 9 维矢量 $x' = (x^T \varphi(x))^T \in R^9$。选择 $\varphi(\ )$ 的最好的方法，是使它成为 x 的模数的函数，它在样本矢量模数值一定范围内是单调地减少的（参阅图 5.3）。

图 5.3 两个一维矢量 x_1 和 x_2 用标量 ϕ 增值为矢量 $(x_1\phi)^T$ 和 $(x_2\phi)^T$. 其模数差已转换为角度差. 其中 x_1', x_2' 为用常数增值的. 矢量 x_1'', x_2'' 为用函数 $\phi=\phi(x)$ 增值的

经过这样的预处理后，再将 LSM 算法应用于具有 9 维矢量的设计集中。用了十个训练循环，把子空间旋转到最佳方向。目前分类的最终精度，依增值函数 $\varphi(.)$ 的不同值，分别为试验数据的 94% 到 97%。作为对照方法的 ML分类器，其精度也差不多或稍好（1—2%）一些，但其速度却慢四倍。

通常，使用学习子空间法意味着在用训练样本设计分类器和对模式进行实际分类时两者计算量之间一定的折衷。Eloranta 对 LSM 法和 ML 分类器（式 5.28）进行过某些比较。n 维模式空间的类别协方差矩阵求逆计算，需进行 n^3 次

• 114 •

b)

Figure 1. Document text image: a) before binarization, b) after binarization

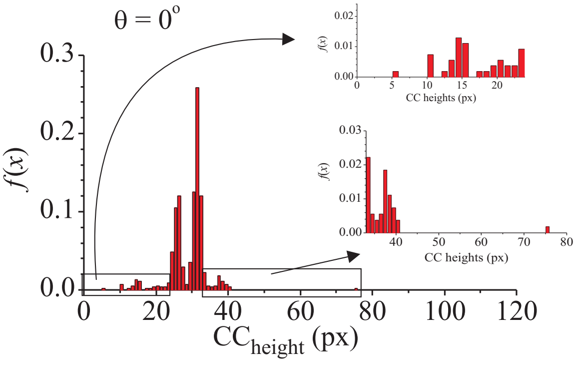
to 1, i.e. $\Delta\mathrm{CC}_{\mathrm{height}} = 1$. Hence, the histograms of relative frequencies correspond to density distributions:

$$f(x) = \frac{N_i}{N\,\Delta\mathrm{CC}_{\mathrm{height}}}, \tag{1}$$

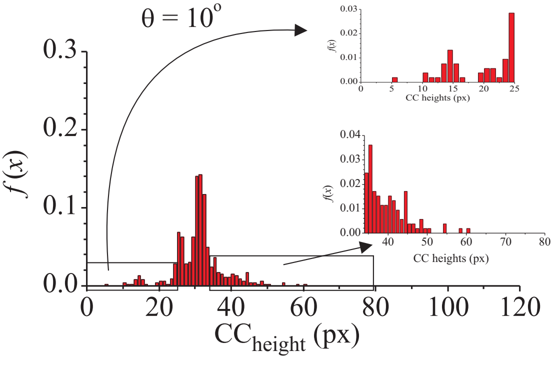
where N_i is the number of data in each class, while N is the total number of data.

These objects are excerpts from document image rotated at different angles, i.e. $0°$, $10°$, $20°$ and $30°$. The histogram of the same document image rotated at different angles are drawn with a different number of data. For example, the histograms are drawn with the following number of data: a) the angle $0°$ with 541 data, b) the angle $10°$ with 528 data, c) the angle $20°$ with 522 data, and d) the angle $30°$ with 522 data. It is a consequence of the object size enlargement established by joining different objects due to the rotation angle. However, histograms are normalized to 1, i.e. to distribution functions. Hence, a different number of data cannot be seen on histograms. Furthermore, the histograms have a considerable number of cavities (class without data) on the left and right tails. In order to clearly present the cavities in histograms, the left and right tails are zoomed and shown as sub-images in Figure 2. This phenomenon is more expressed for larger skew angles. Taking these facts into account makes the document image preprocessing even more difficult.

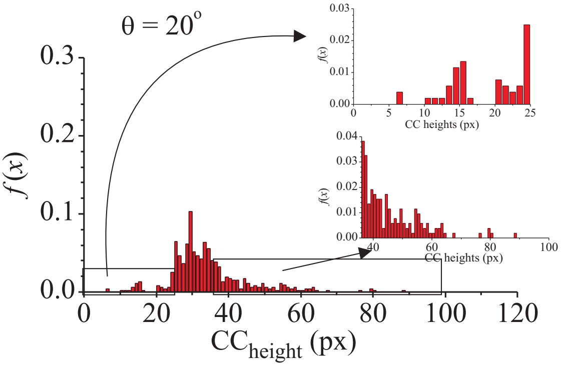
The elements from the tails of the histogram should be excluded because they represent redundant data. These redundant data include small characters, part of characters, noise elements in the left tail, and big characters or objects primarily extended due to text skew by joining connected components from different text lines in the right tail. However, some caution in data extraction is mandatory. For statistical analysis, the number of data elements should be representative. Typically, the number of text elements in the document image representing each page is between


θ = 0°
f(x)
CC_height (px)
CC heights (px)
f(x)
CC heights (px)
f(x)

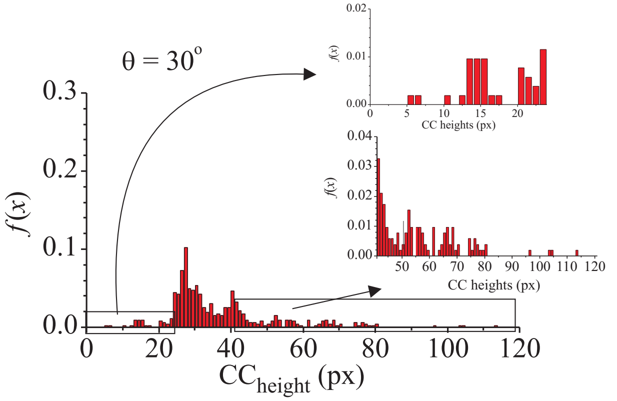
a)


θ = 10°
f(x)
CC_height (px)
CC heights (px)
f(x)
CC heights (px)
f(x)

b)


θ = 20°
f(x)
CC_height (px)
CC heights (px)
f(x)
CC heights (px)
f(x)

c)



d)

Figure 2. Distributions of object heights in document image: a) skew angle 0°, b) skew angle 10°, c) skew angle 20°, d) skew angle 30°

500 and 2 000 elements. Such number of data is representative from the statistical point of view [20]. Furthermore, due to the nature of a printed text, each element incorporates similar or the same text skew attribute. In order to exclude noise or redundant elements, text elements with smallest and largest heights can be freely excluded. However, the right questions is what is the amount of data that should be excluded.

The proposed method is based on data rejection linked with the object heights that do not fall within ranges based on the percentile values of distribution. The percentile of the distribution is the value of a variable below which a certain percentage of observations fall [20]. In order to optimize data rejection process, percentile distribution ranges of 5–95 %, 10–90 %, 15–85 %, 25–75 % are investigated.

To find the optimum percentile distribution range for data reduction, the covariance analysis of two random variables X and Y is made. Let the random variable X represent the difference between maximal and minimal object heights, and the random variables Y represent the difference between the measured and the reference text skew.

Briefly, the variance of a variable is the measure of data dispersion, while the covariance indicates how two variables vary together. Covariance between random variables X (with values $x_1, x_2, \ldots, x_N$) and Y (with values $y_1, y_2, \ldots, y_N$), with same number of values N, is given as [20]:

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\left[(x_i - \overline{x})(y_i - \overline{y})\right]. \tag{2}$$

For uncorrelated values X and Y the covariance is zero. For correlated variables the covariance is different from zero. If the $cov(X,Y) > 0$, then Y will tend to increase as X increases, and if $cov(X, Y) < 0$, then Y will tend to decrease as X increases. Therefore, the minimum value of the absolute covariance between X and Y leads to the smallest output value Y for the full range of input values X.

Figure 3 shows difference height values for different choices of percentile distribution ranges.

It indicates distinctive values in a distribution like minimum heights, maximum heights, mean and median as well as used percentiles for the document image, which is rotated for the full range of angles (up to 30°). Above statistical approach, which uses text elements with their heights is proposed in decision making process of detecting redundant data. Accordingly, the color part of images represents data which are included in further processing (see Figure 3).

Figure 4 shows the exclusion of redundant data given in the text image.

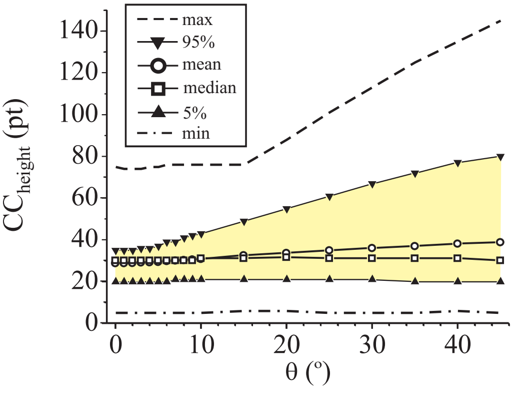
2.3 Text Skew Estimation Based on the Extended Convex Hulls

The algorithm for the text skew estimation, which is proposed in [13, 21], exploited the bounding boxes to define and extract the text regions. It can be said that they are the prerequisite for the connected component creation. In contrast, the convex hull [22] is the smallest convex polygon that contains all its points. Hence, the convex hull includes a smaller region than the bounding box [15]. This circumstance is shown in Figure 5.

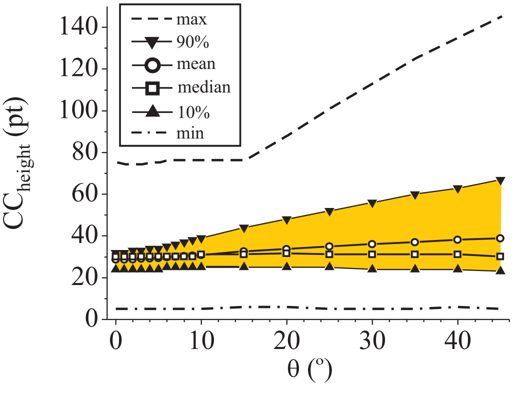
As can be seen from Figure 5, the fragments of touching neighbor text's object could be reduced by using convex hulls instead of bounding boxes. Accordingly, we decided to use the above approach in the algorithm.

The algorithm starts when the convex hulls are designated around the text objects in complementary document image. They assign the regions that will be filled. After filling process, they establish closed regions. Currently, some of the closed regions are joined. They create short connected components (CC). Such a text image is given with matrix $\mathbf{C}$. To correctly estimate the text skew, connected components should be extended in order to emulate a full or partial text line. Hence, morphological dilation is applied to the complementary image $\mathbf{C}$. This way, the adjacent CC's are merged establishing parts of the text line. Structuring element $\mathbf{S}$ representing a horizontal line with variable width is used. In order not to touch or join separate neighbor text lines, the width of the line should be chosen carefully. It heavily depends on height of each CC. Empirically, it is given as approximately 30 % of the connected component's height. Furthermore, it should be noted that suppression of redundant data will exclude the connected components with the smallest and the biggest heights (part on the left and the right tail from histogram) before the application of morphological operators (see Figure 6 d) for reference). This way, the probability of merging the connected components from neighbor text lines is minimized. Morphological operation is given as:

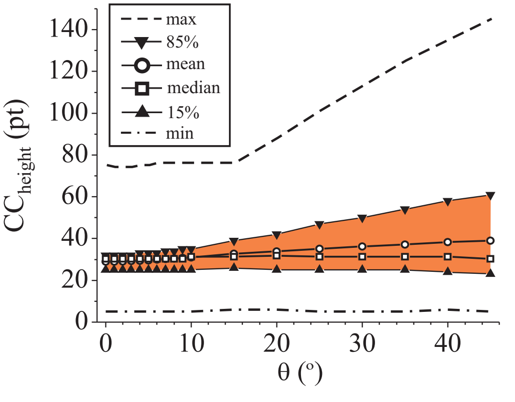
$$\mathbf{Y} = \mathbf{C} \oplus \mathbf{S}. \tag{3}$$


max
95%
mean
median
5%
min
140
120
100
80
60
40
20
0
0
10
20
30
40
CC_{height} (pt)
θ (°)

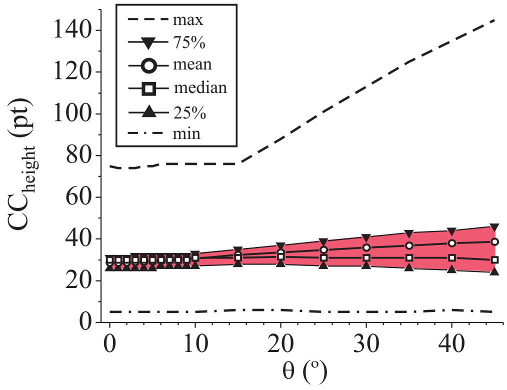
a)


max
90%
mean
median
10%
min
140
120
100
80
60
40
20
0
0
10
20
30
40
CC_{height} (pt)
θ (°)

b)

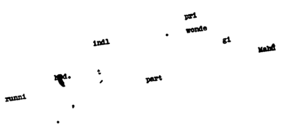

max
85%
mean
median
15%
min
140
120
100
80
60
40
20
0
0
10
20
30
40
CC_{height} (pt)
θ (°)

c)

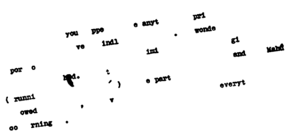


d)

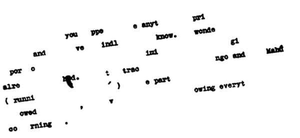
Figure 3. Difference height values inclusion for different ranges of percentile distribution: a) percentiles 5–95, b) percentiles 10–90, c) percentiles 15–85, d) percentiles 25–75



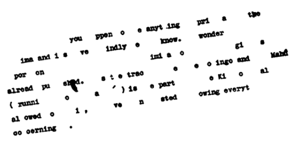
a)



b)



(c)



d)

Figure 4. Redundant data excluded with a proposed statistical approach for the text skew of 8°: a) percentile range 5–95, b) percentile range 10–90, c) percentile range 15–85, d) percentile range 25–75

It is clear that the longest connected component in **Y** is characterized with the orientation similar to dominant text skew [13, 21]. In order to identify dominant text skew, the longest connected components should be extracted. To find it, Euclidean distance between the most bottom point $p(x_p, y_p)$ and the most top point $q(x_q, y_q)$ of each connected component CC is calculated. The longest connected component CC_L is extracted by finding the connected component with the longest Euclidean distance:

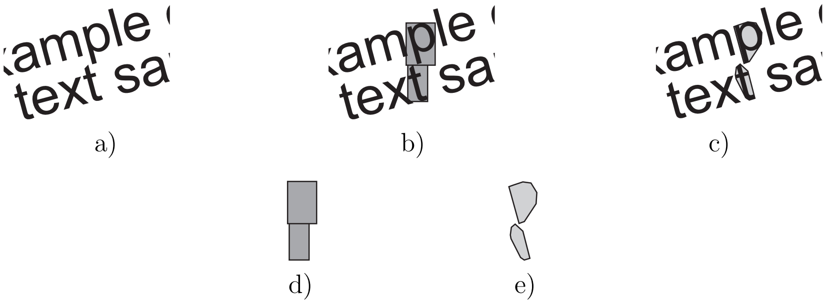


Figure 5. Comparison between connected components and convex hulls: a) fragment of the text, b) fragment of the text with extracted bounding boxes, c) fragment of the text with extracted convex hulls, d) extracted bounding boxes, e) extracted convex hulls

$$\mathrm{CC}_L = \underbrace{\max}_{k} \left(\sqrt{(x_p^k - x_q^k)^2 + (y_p^k - y_q^k)^2} \right), \tag{4}$$

where $k = 1, \ldots, K$ is the total number of connected components.

Furthermore, the orientation of the longest connected component CC_L should be identified by an appropriate method like the moments. Moment defines the measure of the pixel distribution in the image. It depends on the object contour. Moments of the binary image $\mathbf{Y}$ featuring M rows and N columns are [11]:

$$m_{pq} = \sum_{i=1}^{M} \sum_{j=1}^{N} i^p j^q, \tag{5}$$

where p and $q = 0, 1, 2, 3, \ldots, r$, and r represents the order of the moment. The central moments μ_{pq} of the binary image can be calculated as:

$$\mu_{pq} = \sum_{i=1}^{M} \sum_{j=1}^{N} (i - \overline{x})^p (j - \overline{y})^q. \tag{6}$$

The image feature which represents the object orientation θ is obtained from the moments. It illustrates the angle between the object and the horizontal axis. This angle is calculated as [11]:

$$\theta = \frac{1}{2} \arctan \left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}} \right). \tag{7}$$

Hence, the orientation of the longest connected component CC_L can estimate the global text skew θ. According to the orientation θ, the initial document image is de-skewed.

Figure 6 shows all steps of the proposed algorithm for text skew estimation.

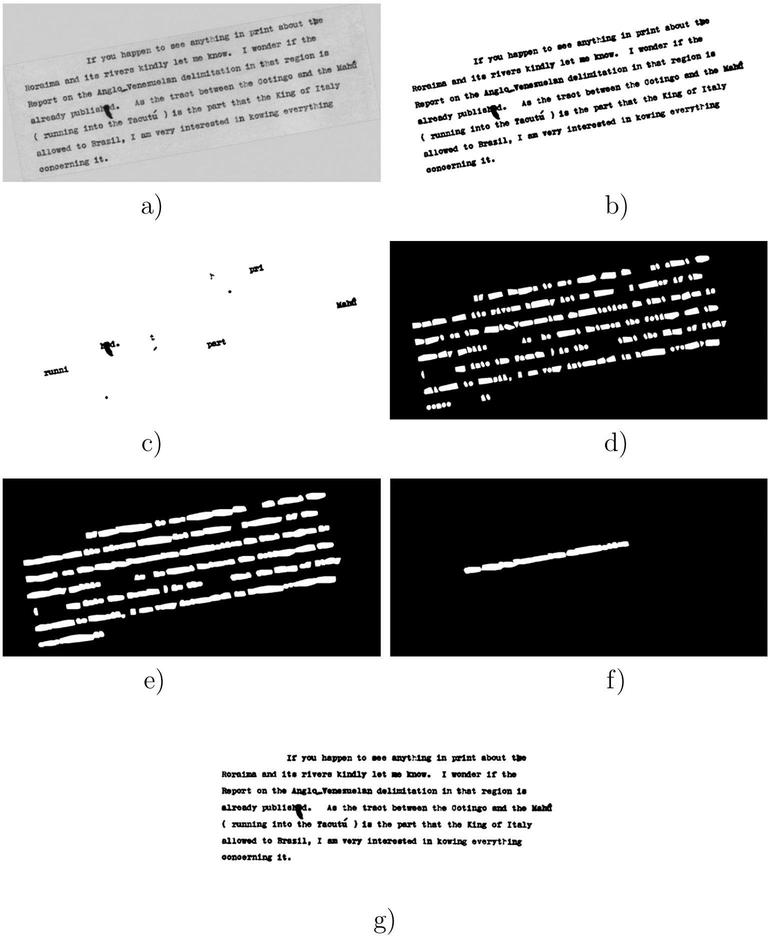


Figure 6. Different steps of the text skew algorithm: a) initial document, b) binarized document, c) preprocessing exclusion, d) convex hull extraction, e) CC extension with morphology, f) longest CC extraction, g) document de-skew

3 EXPERIMENTS

The main goal of the experiment is the evaluation of the algorithm's ability to estimate a text skew. It is performed on real datasets. Dataset samples are rotated for the angle θ, from $0°$ to $10°$ by $1°$ and from $10°$ to $40°$ by $5°$ steps around the origin. It is shown in Figure 7.

Dataset consists of document images in Latin, Cyrillic and Chinese. Some of them represent document samples that are older than 100 years. The samples are obtained by scanning at the resolution of 300 dpi. Figure 8 illustrates samples from the dataset.

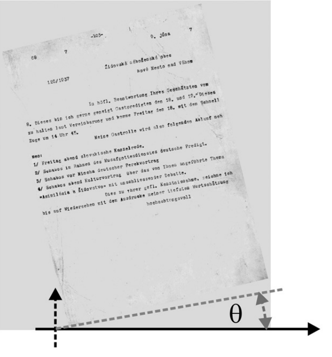


Figure 7. Document image rotation

The results obtained after the algorithm's application to dataset samples are evaluated by the absolute deviation:

$$\Delta\theta_A = |\theta_{REF} - \theta_A|, \tag{8}$$

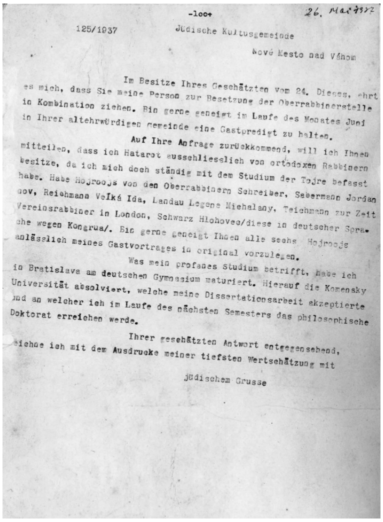
where θ_{REF} is the reference skew of the input text sample and θ_A is text skew estimated by the algorithm. Reference skew is obtained by rotating the image around origin for given angles. However, some of the samples from database are parts of Document Image Skew Estimation Contest – ICDAR 2013 with previously defined skew angles.

4 RESULTS AND DISCUSSION

Measured results are obtained for different percentiles range in the preprocessing stage of the algorithm. These results, in a form of absolute deviation for the full range of rotation angles ($0° - 40°$) applied to document image from dataset, is shown in Table 1. They can be classified as:

- for data excerpt from percentile range 5-95 the absolute deviation is from 0.0221° to 0.7996° with average value of 0.3786°,
- for data excerpt from percentile range 10-90 the absolute deviation is from 0.0115° to 0.4988° with average value of 0.1863°,
- for data excerpt from percentile range 15-85 the absolute deviation is from 0.0221° to 0.7471° with average value of 0.3131°,
- for data excerpt from percentile range 25-75 the absolute deviation is from 0.0221° to 2.8414° with average value of 1.1733°.

Figure 9 shows $\mathrm{cov}(X, Y)$ as a function of the percentile distribution range obtaining from the results of our experiment. Because of the choice of X and Y, if the



-100+ 26. Mai 1937

125/1937 Jüdische Kultusgemeinde

Nové Mesto nad Váhom

Im Besitze Ihres Geschätzten vom 24. Dieses, ehrt es mich, dass Sie meine Person zur Besetzung der Oberrabbinerstelle in Kombination ziehen. Bin gerne geneigt im Laufe des Monates Juni in Ihrer altehrwürdigen Gemeinde eine Gastpredigt zu halten.

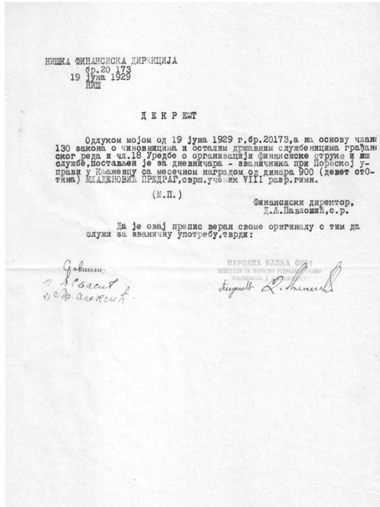
Auf Ihre Anfrage zurückkommend, will ich Ihnen mitteilen, dass ich Hatarot ausschliesslich von ortodoxen Rabbinern besitze, da ich mich doch ständig mit dem Studium der Tojre befasst habe. Habe Hojroojs von den Oberrabbinern Schreiber, Sabermann Jordanov, Reichmann Veľká Ida, Landau Legene Michalany, Teichmann zur Zeit Vereinsrabbiner in London, Schwarz Hlohovec/diese in deutscher Sprache wegen Kongrua/. Bin gerne geneigt Ihnen alle sechs Hojroojs anlässlich meines Gastvortrages in original vorzulegen.

Was mein profanes Studium betrifft, habe ich in Bratislava am deutschen Gymnasium maturiert. Hierauf die Komensky Universität absolviert, welche meine Dissertationsarbeit akzeptierte und an welcher ich im Laufe des nächsten Semesters das philosophische Doktorat erreichen werde.

Ihrer geschätzten Antwort entgegensehend, zeichne ich mit dem Ausdrucke meiner tiefsten Wertschätzung mit

jüdischem Grusse

a)



НИШКА ФИНАНСИСКА ДИРЕКЦИЈА
бр.20 173
19 јуна 1929
НИШ

ДЕКРЕТ

Одлуком мојом од 19 јуна 1929 г,бр.20173,а на основу члана 130 закона о чиновницима и осталим државним службеницима грађанског реда и чл.18 Уредбе о организацији финансиске струке и ... службе,постављен је за дневничара - званичника при Пореској управи у Књажевцу са месечном наградом од динара 900 (девет стотина) МЛАДЕНОВИЋ ПРЕДРАГ,сврш.ученик VIII разр.гимн.

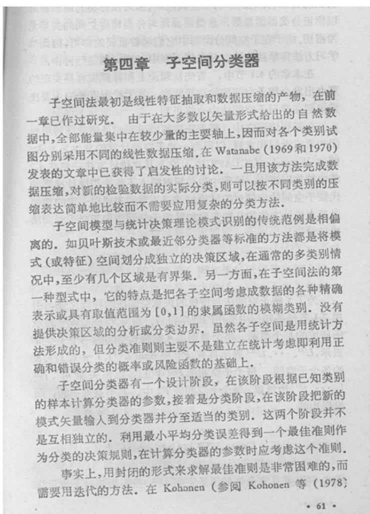
(М.П.)

Финансиски директор,
Д.Љ.Павловић,с.р.

Да је овај препис веран своме оригиналу с тим да служи за званичну употребу,тврди:

НАРОДНА БАНКА ФНРЈ
ЦЕНТРАЛА ЗА НАРОДНУ РЕПУБЛИКУ СРБИЈУ
ФИЛИЈАЛА У КЊАЖЕВЦУ

b)



第四章　子空间分类器

子空间法最初是线性特征抽取和数据压缩的产物，在前一章已作过研究．由于在大多数以矢量形式给出的自然数据中，全部能量集中在较少量的主要轴上，因而对各个类别试图分别采用不同的线性数据压缩．在 Watanabe（1969和1970）发表的文章中已获得了启发性的讨论．一旦用该方法完成数据压缩，对新的检验数据的实际分类，则可以按不同类别的压缩表达简单地比较而不需要应用复杂的分类方法．

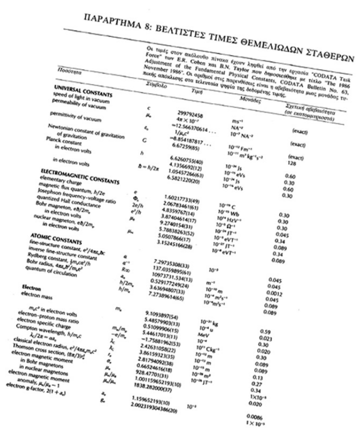
子空间模型与统计决策理论模式识别的传统范例是相偏离的．如贝叶斯技术或最近邻分类器等标准的方法都是将模式（或特征）空间划分成独立的决策区域，在通常的多类别情况中，至少有几个区域是有界集．另一方面，在子空间法的第一种型式中，它的特点是把各子空间考虑成数据的各种精确表示或具有取值范围为［0,1］的隶属函数的模糊类别．没有提供决策区域的分析或分类边界．虽然各子空间是用统计方法形成的，但分类准则则主要不是建立在统计考虑即利用正确和错误分类的概率或风险函数的基础上．

子空间分类器有一个设计阶段，在该阶段根据已知类别的样本计算分类器的参数，接着是分类阶段，在该阶段把新的模式矢量输入到分类器并分至适当的类别．这两个阶段并不是互相独立的．利用最小平均分类误差得到一个最佳准则作为分类的决策规则，在计算分类器的参数时应考虑这个准则．

事实上，用封闭的形式来求解最佳准则是非常困难的，而需要用迭代的方法．在 Kohonen（参阅 Kohonen 等（1978）

· 61 ·

c)



ΠΑΡΑΡΤΗΜΑ 8: ΒΕΛΤΙΣΤΕΣ ΤΙΜΕΣ ΘΕΜΕΛΙΩΔΩΝ ΣΤΑΘΕΡΩΝ

Οι τιμές στον ακόλουθο πίνακα έχουν ληφθεί από την εργασία "CODATA Task Force" των E.R. Cohen και B.N. Taylor που δημοσιεύθηκε με τίτλο "The 1986 Adjustment of the Fundamental Physical Constants, CODATA Bulletin No. 63, November 1986". Οι αριθμοί στις παρενθέσεις είναι η αβεβαιότητα μιας μονάδας τυπικής απόκλισης στα τελευταία ψηφία της δεδομένης τιμής.

d)

Figure 8. Samples from dataset: a) Latin document, b) Serbian Cyrillic document, c) Chinese document, d) Greek Cyrillic document

Percentiles	5–95	10–90	15–85	25–75
$\theta_{REF}(^\circ)$		$\Delta\theta$		
0	0.1055	0.1031	0.6296	0.2531
1	0.1006	0.3074	0.6840	0.0290
2	0.7337	0.0634	0.1067	0.1067
3	0.7371	0.0260	0.1587	0.1587
4	0.7871	0.0115	0.0239	0.2157
5	0.2458	0.0437	0.2458	0.2458
6	0.0221	0.2362	0.0221	0.0221
7	0.2889	0.2732	0.2889	0.2889
8	0.7996	0.0551	0.0504	0.0504
9	0.7842	0.2182	0.0987	2.3029
10	0.2775	0.0724	0.2775	2.1543
15	0.3697	0.2931	0.2939	2.6840
20	0.3497	0.2984	0.5437	2.8414
25	0.2223	0.4988	0.3462	1.7993
30	0.2245	0.2410	0.7471	2.5398
35	0.1856	0.2147	0.3882	2.0709
40	0.2028	0.2114	0.4171	2.1832
Average (0°–40°)	0.3786	**0.1863**	0.3131	1.1733

Table 1. Absolute deviation $\Delta\theta$ vs. percentiles range choice

covariance has the lowest value, then the estimated text skew deviation will be the smallest. According to the experiment, the best results are obtained for 10–90 % percentile, which uses 80 % of measured data around the mean value (see Table 1). This way, it is shown that decision of choosing the relevant text data from aforementioned percentile range is the right choice of the algorithm for text skew estimation which is based on boundary growing area. Hence, the proposed methodology can be used for excluding redundant text elements in each document image. Consequently, it can be used as a standard methodology for disregarding the redundant data using such types of the algorithm.

The proposed algorithm has the average absolute deviation of 0.1863° for the text skew angle θ up to 40°. This result is slightly better than the value of 0.1934 which is obtained in the experiment applied to the same database for the algorithm proposed in [16]. Furthermore, Chou's algorithm is supposed to be used for the skew angle range 0°–15° only. Therefore, our algorithm has quite acceptable values of the absolute deviation in the wide range of angles. Typically, the absolute deviation of 0.1°−0.2° for the text skew angle range of up to 15° is considered to be good results [16, 17]. Furthermore, the algorithm has been applied to different types of documents (including a part of Document Image Skew Estimation Contest – ICDAR 2013) and different types of letters. It can be used for detecting skew in documents like letters, technical articles, journals, dictionary, project documentations. It shows a correctness and robustness in various occasions. However, it cannot identify the

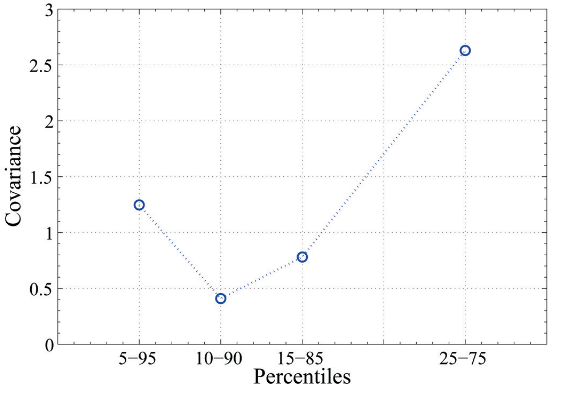


Figure 9. Absolute covariance vs. percentiles choice for the full range of angles (0°– 40°)

text skew immanent to comics, especially if the text in pictures is not relevant as an input to determine the text skew. Furthermore, it does not have the versatility of multi-stage methods. These methods include complex procedures of geometrical filtering in preprocessing stages in order to exclude redundant data [5, 6]. However, such methods are much more computer time intensive.

In further developing steps, proposed method should be expanded with the inclusion of some additional geometrical filtering steps. These steps will contribute to lower dispersion of absolute error value (up to 0.1°).

5 CONCLUSIONS

The paper proposed the robust method for the estimation of the global text skew. It was based on the convex hull extraction over the text in a document image. Furthermore, convex hulls were extended with binary morphology. After that, the longest object was extracted. On the basis of its contour, the moment based method estimated its orientation, which represents the global text skew. The method was examined on the real dataset representing document images in Latin, Cyrillic, Greek and Chinese. It showed good results in estimating the global text skew of document images given in the resolution of 300 dpi. The obtained results are quite promising. Further improvement of the method will be made toward the inclusion of the additional geometrical filtering in the preprocessing stage, which is supposed to lower the dispersion of the absolute error value.

Acknowledgement

This work was partially supported by the Grant of the Ministry of Science from Republic of Serbia, as a part of the project TR33037 and III43011 within the framework of Technological development program. The third author is supported by National Natural Science Foundation of China (No. 61261130590) and 973 National Basic Research Program of China (No. 2014CB340506).

REFERENCES

[1] Wen, P. C.—Wang, L. L.: Identification of Technical Journals by Image Processing Techniques. Computing and Informatics, Vol. 29, 2010, No. 2, pp. 165–182.

[2] Amin, A.—Wu, S.: Robust Skew Detection in Mixed Text/Graphics Documents. Proceedings of 8th International Conference on Document Analysis and Recognition (ICDAR), Seoul, Korea, 2005, pp. 247–251.

[3] Manmatha, R.—Srimal, N.: Scale Space Technique for Word Segmentation in Handwritten Manuscripts. In: M. Nielsen et al. (Eds.): Scale-Space Theories in Computer Vision, Proceedings of the Second International Conference Scale-Space Theories in Computer Vision, Corfu, Greece, 1999, LNCS, Vol. 1682, 1999, pp. 22–33.

[4] O'Gorman, L.: The Document Spectrum for Page Layout Analysis. IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 15, 1993, No. 11, pp. 1162–1173.

[5] Saragiotis, P.—Papamarkos, N.: Local Skew Correction in Documents. International Journal of Pattern Recognition and Artificial Intelligence, Vol. 22, 2008, No. 4, pp. 691–710.

[6] Makridis, M.—Nikolau, N.—Papamarkos, N.: An Adaptive Technique for Global and Local Skew Correction in Color Documents. Expert Systems and Applications, Vol. 37, 2010, No. 10, pp. 6832–6843.

[7] Zagoris, K.—Chatzichristofis, S. A.—Papamarkos, N.: Text Localization Using Standard Deviation Analysis of Structure Elements and Support Vector Machines. EURASIP Journal on Advances in Signal Processing, 2011, No. 47, pp. 1–12.

[8] Das, A. K.—Chanda, B.: A Fast Algorithm for Skew Detection of Document Images. International Journal on Document Analysis and Recognition, Vol. 4, 2001, No. 2, pp. 109–114.

[9] Najman, L. A.: Using Mathematical Morphology for Document Skew Estimation. In: B. Smith et al. (Eds.): Document Recognition and Retrieval XI, Proceedings of the SPIE, Vol. 5296, 2004, pp. 182–191.

[10] Nguyen, D. T.—Vo, D. B.—Nguyen, T. M.—Nguyen, T. G.: A Robust Document Skew Estimation Algorithm Using Mathematical Morphology. Proceedings of 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007), Paris, France, 2007, Vol. 1, pp. 496–503.

[11] Kapogiannopoulos, G.—Kalouptsidis, N.: A Fast High Precision Algorithm for the Estimation of Skew Angle Using Moments. Proceedings of International Conference on Signal Processing, Pattern Recognition, and Applications, Crete, Greece, 2002, pp. 275–279.

[12] Manjunath Aradhya, V. N.—Hemantha Kumar, G.—Shivakumara, P.: Skew Estimation Technique for Binary Document Images Based on Thinning and Moments. Journal of Engineering Letters, Vol. 14, 2007, No. 1, pp. 127–134.

[13] Brodić, D.—Milivojević, D. R.: An Algorithm for the Estimation of the Initial Text Skew. Information Technology and Control, Vol. 41, 2012, No. 3, pp. 211–219.

[14] Brodić, D.—Milivojević, Z. N.: Handwritten Text Skew Based on Binary Moment. Radioengineering, Vol. 21, 2012, No. 1, pp. 162–169.

[15] Brodić, D.—Peng, L.—Maluckov, Č.—Milivojević, Z. N.: Detection of the Global Text Skew with Initial Skew Rate. Proceedings of 36th International Conference on Telecommunications and Signal Processing (TSP 2013), Rome, Italy, 2013, pp. 853–857.

[16] Chou, C.-H.—Chu, S.-Y.—Chang, F.: Estimation of Skew Angles for Scanned Documents Based on Piecewise Covering by Parallelograms. Pattern Recognition, Vol. 40, 2007, No. 2, pp. 443–455.

[17] Mascaro, A. A.—Cavalcanti, G. D. C.—Mello, C. A. B.: Fast and Robust Skew Estimation of Scanned Documents Through Background Area Information. Pattern Recognition Letters, Vol. 31, 2010, No. 11, pp. 1403–1411.

[18] Otsu, N.: A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, 1979, No. 1, pp. 62–66.

[19] Chen, K.-N.—Chen, C.-H.—Chang, C.-C.: Efficient Illumination Compensation Techniques for Text Images. Digital Signal Processing, Vol. 22, 2012, No. 5, pp. 726–733.

[20] Wackerly, D. D.—Mendenhall, W. III.—Scheaffer, R. L.: Mathematical Statistics with Applications. Duxbury Press, Belmont, U.S.A., 1996.

[21] Brodić, D.: The Evaluation of the Initial Skew Rate for Printed Text. Journal of Electrical Engineering, Vol. 62, 2011, No. 3, pp. 142–148.

[22] Kim, Ch. E.—Stojmenovic, I.: Sequential and Parallel Approximate Convex Hull Algorithms. Computing and Informatics, Vol. 14, 1995, No. 6, pp. 597–610.

Darko Brodić received his B.E.E. and M.E.E. from the Faculty of Electrical Engineering, University of Sarajevo in 1987 and 1990, as well as Ph.D. in electrical engineering from the Faculty of Electrical Engineering, University of Banja Luka in 2011. Now he is Assistant Professor of computer science at the Technical Faculty in Bor, University of Belgrade, Serbia. His current research interests include different aspects of signal processing, natural image processing, pattern recognition and document image processing.

Čedomir A. Maluckov received his B.Sc. and M.Sc. degrees in physics and Ph.D. degree in electrical engineering from the Faculty of Electronics, University of Niš, Serbia, in 1996, 2002, and 2004, respectively. He is currently Associate Professor of physics at the Technical Faculty in Bor, University in Belgrade, Serbia. His research interests include stochastic modeling, physical description of electrical discharge in gases and applied statistics in pattern recognition.

Liangrui Peng received her B.E.E. and M.E.E. from the Huazhong University of Science and Technology, Wuhan, Hubei, China in 1994 and 1997, as well as Ph.D. from the Tsinghua University, Beijing, China in 2010. She is currently Associate Professor at the Department of Electronic Engineering, Tsinghua University, Beijing, China. Her research interests include pattern recognition, image and document information processing.