UDC 004.932.7

O. Gorokhovatskyi, Ph.D.

CLASSIFICATION OF TEXT AND GRAPHIC REGIONS ON A FRAGMENT OF DOCUMENT IMAGE

Abstract. *Paper describes method for graphic and text image region classification. Construction of the narrative function characterizing areas of the image sub-elements is proposed. This function can be used for training and classification using support vector machines. Experimental analysis of the method's quality showed that correct classification can be performed with probability greater than 0.94, it also allowed to identify scope and limitations of the method.*

Keywords: *image classification, document region, threshold, decomposition, text region, graphic region, areas distribution, support vector machine*

А. В. Гороховатский, канд. техн. наук

КЛАССИФИКАЦИЯ ТЕКСТОВЫХ И ГРАФИЧЕСКИХ ОБЛАСТЕЙ НА ФРАГМЕНТЕ ИЗОБРАЖЕНИЯ ДОКУМЕНТА

Аннотация. *Статья посвящена разработке метода классификации графической и текстовой области изображения. Предложено построение описательной функции, характеризующей размеры объектов на изображении, которая может быть использована для обучения и классификации с помощью машины опорных векторов. Проведенные экспериментальные исследования качества работы метода показали возможность выполнять корректную классификацию с вероятностью более 0.94 и позволили выделить области применения и ограничения метода.*

Ключевые слова: *классификация изображения, область документа, порог, разбиение, текстовая область, графическая область, распределение площадей, машина опорных векторов*

О. В. Гороховатський, канд. техн. наук

КЛАСИФІКАЦІЯ ТЕКСТОВИХ ТА ГРАФІЧНИХ ОБЛАСТЕЙ НА ФРАГМЕНТІ ЗОБРАЖЕННЯ ДОКУМЕНТА

Анотація. *Статтю присвячено розробці методу класифікації графічної і текстової області зображення. Запропоновано побудову описової функції, що характеризує розміри об'єктів на зображенні, яка може бути використана для навчання та класифікації за допомогою машини опорних векторів. Проведені експериментальні дослідження якості роботи методу показали можливість виконувати коректну класифікацію з імовірністю більше 0.94 і дозволили визначити області застосування і обмеження методу.*

Ключові слова: *класифікація зображення, область документу, поріг, декомпозиція, текстова область, графічна область, розподілення площ, машина опорних векторів*

Introduction

Automatic analysis of electronic documents in modern computerized world has long been an everyday process that involves a huge number of people. Digitizing documentation for electronic processing involves multiple steps, such as scanning, segmentation, analysis, recognition, etc. During this processing work with a document as an image is unavoidable, since its transformation to an image (e.g., using photography) is a typical method of transferring such information.

During the automatic analysis of a document image situation of uncertainty often occurs, which requires adoption of different decisions (by human or algorithmically) depending on the type of the content being processed. Often such analysis is used for the processing content of homogeneous content type, such as text or graphic region on the image.

To accomplish this task ability to recognize content type during the image fragment processing is required, content type is typically represented by label "text" or "not a text". In turn, during solving real problems label "not a text" often assigned to a graphic object (such as chart, illustration, drawing, figure, etc.) or a table object on the image. This paper focuses on the investigation of the classification of text and graphic regions as one of stages for document image processing.

Separation of text and graphic regions during document image analysis is not a trivial task, which requires an individual approach to each type of documents. There are two main classes of methods for solving such a problem, the first of which operates with mainly quantitative characteristics of the region, such as density of features, geometric proportions of sizes [5, 8, 12], etc. Main

weakness of these methods is relatively low accuracy, because of the use of features strongly dependent on image and color distribution.

The second class of existing methods [2, 3, 4, 6] uses more complicated methods and is usually applied to the entire document image but and to its segmented parts. Allowing to achieve greater accuracy, they rely on the analytical analysis of an image, analysis of the connected components, morphological operations, usage of different a priori information about possible differences between text and graphic regions. Furthermore, there is a necessity of the parameters adjustment for method adaptation to each particular document, for example, to process images with different resolution.

However, there are practical tasks of graphic and text region classification (such as, for example, recognition, document structure analysis, automatic processing of pictures, text search, etc.) that do not require use of multi-step methods for full image processing, and cannot be effectively solved only by the set of quantitative characteristics.

Paper describes methods which allows to classify segmented region of document image containing a homogeneous content as text or graphic region and can be easily adapted to different types of images.

Basic requirements for the image region

By the text region will denote region of an image that contains only text (meaning dark text on a light background). Graphic region can be represented by color, grayscale or binary image, graphics, drawings, diagrams, etc. Input of the method is a preliminary segmented rectangular fragment of the image document, which requires classification of content type.

Images of text regions are characterized by the structural composition of letters that cannot be said in general about graphic regions, which have more of a random distribution. Moreover, for the text region it is typical to have large number of small segments of similar size corresponding to the location of the letters.

In turn, graphic regions are more sensitive to binarization procedure, because they have larger variety of color components. This fact allows to choose a fixed threshold for binarization, as there is no need to perform an effective segmentation of the graphics region to isolate its significant objects.

The main idea of this paper is the sequential segmentation of the initial region under the assumption that it is a text region, construction of size distribution function for parts and components of the region and classification of this function using support vector machines.

Size distribution function

Binarization is one of the most common preliminary procedure for a lot of computer vision and image processing methods. The main step of the implementation is the choice of the binarization threshold δ, which must be capable of segmenting object of interest on an image. In the context of the problem being solved such an object can be either text (if the text region is considered) or an arbitrary element of graphics region.

Assuming traditional situation with a dark text on a light background threshold δ will be a specific feature, which does not allow to distort structure of the letters on the image but at same time permits to transform significant part of the graphics region to black color, because graphic element often has a wider color range in comparison to the textual one. Experimental research has shown that threshold limit set up to the fixed value $\delta = 180$ is effective enough.

Image I of $m \times n$ pixels is proposed to be considered a priori as a text region. Splitting of the image I horizontally and vertically using traditional projection analysis [1, 10], forming this way list of N individual segments $S = \{s_1, s_2, ..., s_N\}$, where s_i is the area of *i*-th segment, N is the total number of segments potentially describing letters for specific image I. Under the "area" in this context we mean the geometric area of the rectangular segment without analysis of the internal content.

Set of values $\{s_1, s_2, ..., s_N\}$ is normalized in a way: $s_i' = Ls_i / (m \times n)$ with further quantization to L equivalent levels that allows to construct a description of the image region $\{k_1, k_2, ..., k_L\}$, where k_i is a count of normalized segments s_i' belonging to a level i. During update of k_i value for the i-th level all previous levels from 0 to $i-1$ are penalized with an addition of specific value, because the involvement of greater objects is typical for graphic regions. As the amount of penalty value $j \times 3^i$, j=0,...,i-1 was chosen. Thus, penalization procedure allows to expand separation bor-

der between descriptions for text and graphic regions.

Since hitting of segment size into each level has the same probability, normalization of $\{k_1, k_2, ..., k_L\}$ descriptions is also performed followed by obtaining structure-description in the form $\{k'_1, k'_2, ..., k'_L\}$, where $k'_i = k_i / \max_i(k_i)$.

Fig. 1 shows an example of text region, as well as histogram of normalized $\{k'_1, k'_2, ..., k'_L\}$ values with $L = 10$, corresponding digital values are $\{1, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$. Similar size distribution is an etalon for text region as a fragment of an image that contains only valuable objects of insignificant size.

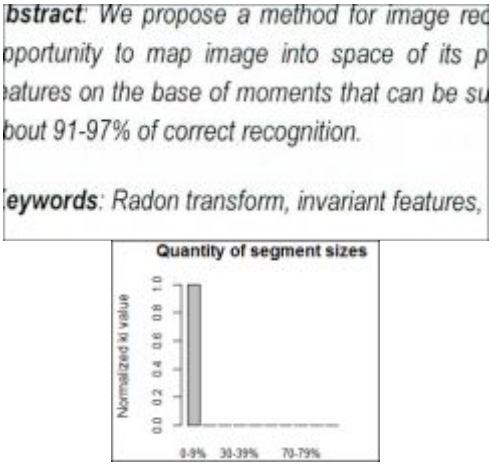
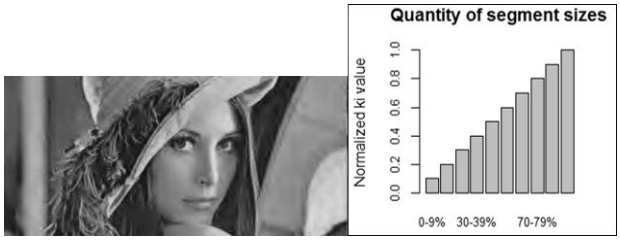


Fig. 1. Example of text region and histogram of normalized values of size distribution

Examples of corresponding values for graphic regions are shown on Fig. 2. As samples of graphic regions we used the full-background Lena image and image with colorful curves on a white background [12], which are for sure not the text regions. As can be seen, the size distribution histogram is significantly different from the one on Fig.1, which makes it possible to classify text and graphic regions. One can also notice, that both histograms are similar despite of the significant difference between the sample images.



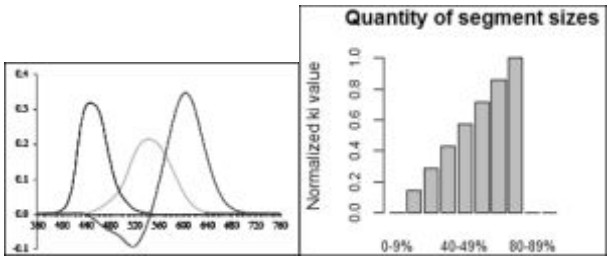


Fig. 2. Examples of graphic regions and corresponding histograms of normalized values of size distribution

Classification

Consider the use of SVM method as a decision-making mechanism to find a solution of binary classification problem in condition of linear inseparability.

Learning and analysis is based on the kernel radial basis function RBF, which can be described as:

$$K(x, x') = (\phi(x), \phi(x')) = \exp(-\gamma \|x - x'\|^2),$$

where x, x' are feature vectors for comparison, $\gamma = -1/2\sigma^2$, σ is the scattering parameter of the kernel, $\|x, x'\|$ is the norm between x, x' vectors. As a norm the simple difference between vectors was used during experimental modelling.

Effectiveness of learning and classification is determined by parameters γ and C – penalty factor which determines the possible error level [11]. Wrong selection of C coefficient can lead to both too high level of learning errors, and too strict conditions of training and classification [11].

Thus, the overall learning model can be written as:

$$\arg\min_{w,\xi,b}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{M}\xi_i\right),$$

$$y_i(w\phi(x_i) - b) \geq 1 - \xi_i,\ 1 \leq i \leq M,$$

where x_i is training vector, y_i is class which supplies training vector, $\xi_i > 0$ is magnitude of error for x_i vector, $w\phi(x) - b = 0$ is separating hyperplane, $w = \sum_{i=1}^{M}\alpha_i y_i \phi(x_i)$, α_i are Lagrange multipliers.

In turn, the classification model for x vector can be represented as:

$$f(x) = sign\left(\sum_{i=1}^{M}\alpha_i y_i K(x, x_i) + b\right).$$

Selection of γ and C parameters is performed as usual by full grid search using values in specified range and choice of combination of parameters providing the minimum error training.

Experimental results

Training set with text regions consists of 1519 images, corresponding set with graphic regions contains 500 images. Limits of C value ranged from 1 to 150 with $\Delta C = 1$ discrete step, corresponding values for γ parameter ranged from 1 to 100 with $\Delta\gamma = 1$ step. After SVM training parameter values $C = 24$, $\gamma = 3$ have been selected as the best ones, error of learning was equal to 0,01955 (less than 2 %), 85 training iterations were performed to achieve such precision.

Quality investigation of the classification method was performed on two different sets of test images, the first of which included 1004 images with text regions (they were generated in half-automatic mode using set of scanned documents LRDE DBD [7]) and 292 graphic regions collected manually. Correct classification of text region can be performed with 0.976 probability, value of 0,8835 can be reached for graphic region. This benchmark was used for detailed analysis of limitations and features of classification method.

Several limitations associated directly with the projection analysis process may occur as a consequence of image decomposition to handle with separate text lines processing. Threshold δ as it was described earlier allows for correct classification of dark text on a light background (but not vice versa), however, a simple preliminary analysis of colors distribution in initial image with its further inversion, if necessary, makes possible elimination of this restriction.

Ignoring deep content analysis in the described approach does not allow to classify the image containing only few large letters (including images that contains single letter) as a text region – in general, this problem can be ambiguous even for manual human analysis. The opposite is also true: if there are several graphic objects that have a relatively small size inside original image fragment, this area is classified as text.

These limitations may be smoothed with presence of specific images with the appropriate properties in training set, either by increasing of quantization levels L count, allowing an adaptation of proposed size distribution set $\{k'_1, k'_2, ..., k'_L\}$ to specific characteristics of images.

Second test set included 4257 images with text regions (derived both from LRDE DBD [7] and Tobacco [9]) and about 15,000 images with graphic regions collected from various documents. Proposed method of classification allows to identifying text region with an accuracy rate of 0.9859 and graphic region with 0.9451.

Fig. 3 shows an image of text region and its size distribution values, which was wrongly classified as graphic because of considerable size of letters. Such errors can be eliminated by extending of training set to include such images.

Images shown on Fig. 4 were erroneously classified as text regions and show typical drawbacks of classification method. First image (Fig. 4) represents dark vertical strips, whose area is small, resulting in values $\{k'_1, k'_2, ..., k'_L\}$ similar to those that correspond to etalon text area (Fig. 1). Another example on Fig. 4 shows graphics region, which contains plurality of small area objects.

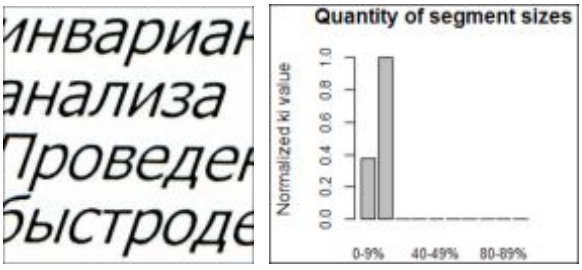


Fig. 3. Example of text region misclassification

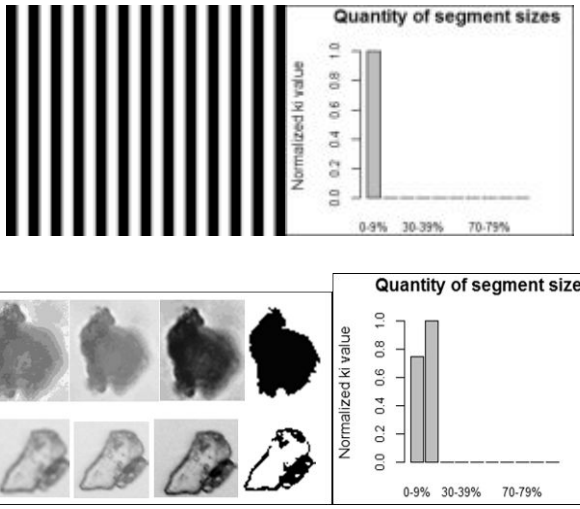


Fig. 4. Examples of incorrect classification of graphic regions

Conclusions

Paper describes the development of the image region classification method, which allows to identify region as text or graphic. Method is based on the construction of descriptive function characterizing the variability of size of objects in image region. Specifics of such function construction implies the use of quantitative values of objects areas that allows to implement this method without significant computational cost. As a classifier the use of support vector machine is proposed, which makes it possible to perform a binary classification (initial image region can be labeled as of two etalon classes – graphics or text region) in general case of linear inseparability of classes. Experimental results showed the effectiveness of classification method which allows to reach probability of 0,98 for text regions and 0,94 for graphic.

The approach we have proposed has some drawbacks and limitations, in particular, the lack of context analysis does not always allow to classify correctly regions with mixed content such as combination of text and graphic elements, as well as areas containing multiple images or text fragments of large size. Elimination of these weaknesses may be the subject of further research and improvement.

References

1. Anupama N., Rupa Ch., and Sreenivasa E. Reddy. Character Segmentation for Telugu Image Document using Multiple Histogram Projections, (2013), *Global Journal of Computer Science and Technology Graphics & Vision*, Vol. 13, Issue 5, Version 1.0 (In English).

2. Bloomberg D.S., Multiresolution Morphological Approach to Document Image Analysis, (1991), *Proceedings of International Conference in Document Analysis and Recognition*, pp. 963 – 971 (In English).

3. Bukhari, S.S., Azawi, M.I.A.A., Shafait, F., and Breuel T.M. Document Image Segmentation using Discriminative Learning over Connected Components, (2010), *Proceeding of the International Workshop on Document Analysis Systems*, ACM New York, NY, USA, pp. 183 – 190 (In English).

4. Bukhari S.S., Shafait F., and Breuel T.M. Improved Document Image Segmentation Algorithm using Multiresolution Morphology, (2011), *Proceedings of the XVIII Document Recognition and Retrieval Conference*, San Jose, CA, USA, January 24 – 29, 2011, pp. 1 – 10 (In English).

5. Gao J., Yang J., Zhang Y., and Waibel A., Text Detection and Translation from Natural Scenes, tech. report CMU-CS-01-139, (2001), *Computer Science Department, Carnegie Mellon University* (In English).

6. Gupta N., and Bange V.K., Image Segmentation for Text Extraction, (2012), *Proceedings of the 2nd International Conference on Electrical, Electronics and Civil Engineering*, Singapore, April 28 – 29, 2012 (In English).

7. Lazzara G., and Geraud T., Efficient Multiscale Sauvola's Binarization, (2013), *International Journal of Document Analysis and Recognition* (In English).

8. Le D.X., Thoma G.R., and Wechsler H. Automated Borders Detection and Adaptive Segmentation for Binary Document Images, (1996), *Proceedings of the International Conference on Pattern Recognition*, Vol. 7276, pp. 737 – 741 (In English).

9. Lewis D., Agam G., Argamon S., Frieder O., Grossman D., and Heard J. Building a Test Collection for Complex Document Information Processing, (2006), *Proceedings of the 29th Annual International ACM SIGIR Conference*, pp. 665 – 666 (In English).

10. Likforman-Sulem L., Zahour A., and Taconet B. Text Line Segmentation of Historical Documents: a Survey, (2007), *International Journal on Document Analysis and Recognition*, Springer, Vol. 9, Issue 2, pp.123 – 138 (In English).

11. SVM Parameters, (2014), available at: http://www.svms.org/parameters/ (accessed 29 January 2014) (In English).

12. Szeliski R. Computer Vision: Algorithms and Applications, Springer-Verlag London Limited, 2011 (In English).

Received 20.02.2014

Gorokhovatskyi Oleksii, PhD, assistant professor of Information and Computer Technics Department, Simon Kuznets Kharkiv National University of Economics, (0572) 702-06-74 add.4-38, Lenin Avenue, 9av, 61166 Ukraine.
E-mail: gorohovatsky@rambler.ru