Document Image Segmentation Based on Wavelet Features

Valery Grishkin

Saint-Petersburg State University
Saint-Petersburg, Russia
e-mail: valery-grishkin@yandex.ru

ABSTRACT

This paper proposes a method for segmentation of images containing both textual and graphical data. The method uses wavelet transformation to build the feature vector and a pattern recognition technique to classify areas of a document image. Values of wavelet coefficients distribution histogram of the source image's sliding window serve as elements of a feature vector. For recognition of document area category, we use a trained classifier based upon support vector machine with RBF kernel.

Keywords

Image processing, document image segmentation, image recognition, wavelet features

1. INTRODUCTION

Nowadays millions of printed documents are scanned to build digital libraries. Classification of these scanned documents would be impossible without image processing and category recognition. This raises a task of dividing a document into a set of separate areas each containing just a text, halftone images, and graphics. Segmentation methods use bottom-up and top-down approaches. A bottom-up approach [1,2,3] attempts to classify relatively small parts of a document that are then joined together and classified again as document areas. These resulting areas are normally of complex shape. A top-down approach [4] first classifies the document as a whole then attempts to segment it into areas of a predefined shape. Algorithms utilizing this approach sometimes fail to process document areas of irregular shape.

This paper proposes a method for segmentation of scanned document images, which is based upon a bottom-up approach. First, the source image is made subject of a multilevel wavelet transformation. Then, small areas of an image are selected using a sliding window. For these small areas, we build wavelet feature vectors, which help us determine the type of each area. Contiguous areas of the same type are joined together to form a region.

2. SEGMENTATION ALGORITHM

2.1. Features extraction

Document areas of different types have textures of different frequency properties. Halftone images have relatively flat textures and relatively uniform spectrum at low frequencies. Textual images and graphics have rapidly changing gradients so their spectrum is biased to higher frequencies. Fourier transform can't calculate spectrum of a small image area with acceptable precision because it divides a space-frequency domain uniformly. In contrast, the wavelet transformation describes frequency properties well enough, regardless of the size of a document area. This multilevel transformation yields a set of scaled images for different frequency ranges with each scaled image being of the same structure as the source image. We propose building the image's feature vector utilizing coefficients of wavelet decomposition. Types of this coefficients distribution law in each frequency range differ among the considered image types. Therefore, the histogram values in each frequency range are used as components of a feature vector.

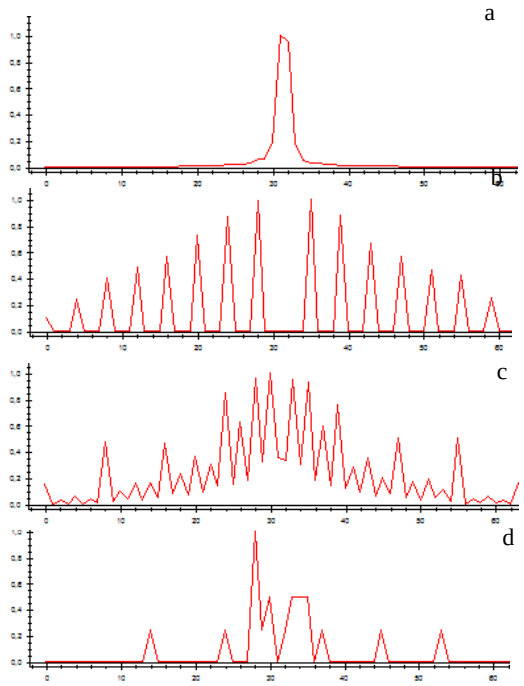


Figure 1: Distribution histograms of wavelet coefficients: (a) photo, (b) text, (c) graphics, (d) background with noise

Figure 1 shows how distribution histograms of wavelet coefficients look like for various image types. These histograms are for X-axis coordinate coefficients of second level wavelet decomposition. The histograms were evaluated from document image fragments of fixed size. Coefficients of zero value weren't included because on this decomposition level there are quite a lot of zeroes for any image type, i.e. zeroes can't be used to distinguish among types. Fragment size should be small enough for fine area localization, and large enough to yield informative features. For calculation of features, a rectangular window of height and width proportional to the height and width of the source image was used. The window was experimentally chosen to be 0.25% square of the source image.

Figure 2 has an algorithm for the evaluation of a feature vector for the image fragment. A 2-level discrete transformation of the source image yields 4 matrices of transformation coefficients on level three: LL2 is a low frequency decomposition matrix, HL2 is a high frequency horizontal decomposition matrix, LH2 is a high frequency

vertical decomposition matrix, and HH2 is a high frequency diagonal decomposition matrix.

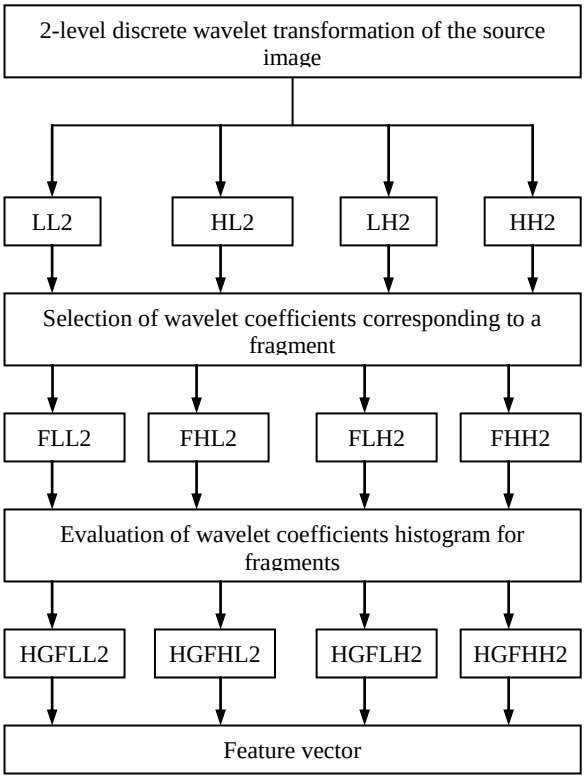


Figure 2: An algorithm for the evaluation of a feature vector for the image fragment

All matrices are 1/16 of the source image in size. From these matrices, submatrices FLL2, FHL2, FLH2, and FHH2, can be extracted for any given image fragment. For each extracted submatrix of coefficients, its distribution histograms HGFLL2, HGFHL2, HGFLH2, and HGFHH2, are built. Zeroes are ignored when building histograms. A sequence of histogram values forms a feature vector. Studies show that best classification results can be achieved with histograms of size 64. This gives us size of a feature vector equal to 256.

2.2. Image fragments classifier

Distinguishing among different types of image fragments is made possible with the help of SVM-based classifier [5]. SVM stands for a support vector machine.

A classifier has been trained against a representative set of scanned and photographed by a camera documents that contained a text, graphics, and halftone images. Prior to classification, areas of interest were pinpointed in every document. Fragments were picked only from these areas of interest. During the training process, five classifiers were built.

Four classifiers were trained to distinguish among two classes each: text T or non-text O (stands for other); halftone image P or non-halftone image O; graphics G or non-graphics O; background B or non-background O.

Fifth classifier was trained to distinguish among four types of images: text T, halftone image P, graphics G (sketches, schemes, drawings), and single-color background with possible noise B.

For training, a cross-validation technique was used. The following SVM kernels were in use: linear kernel, polynomial kernel and an RBF-based kernel (RBF stands for radial basis functions). Of these kernels, RBF kernel demonstrated the best results, that is why we used it in further experimenting.

2.3. Area classification

Trained classifiers are used for the recognition of contiguous document image fragments. Five classifiers will give us five maps of the image with marked (classified) fragments. Maps 1 – 4 have only two classes each and have corresponding fragments marked with 1 and 0. To relieve classification errors, each of these maps is subject to median filtering.

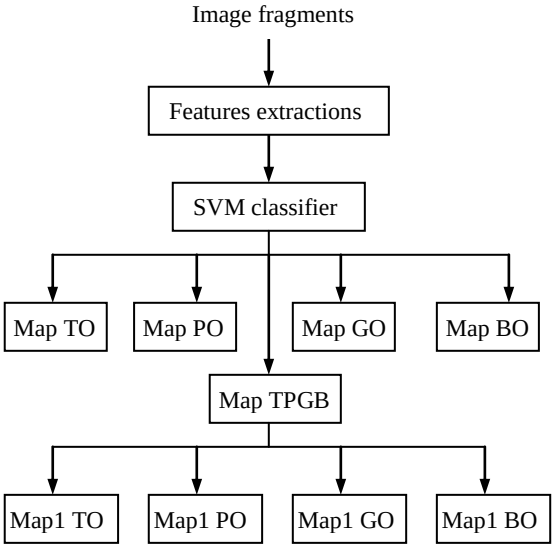


Figure 3: Building maps of classified fragments

In the fifth map, each fragment can be encoded with a number in range of 1 to 4. From these maps, another four maps are derived. Each of these new maps contains only fragments of a given type, encoded as 1, with other fragments zeroed. These additional maps are also subject to a median filter. Building maps of classified fragments is illustrated in Figure 3.

Now we have four pairs of maps of classified fragments: a pair of maps of text fragments distribution Map TO – Map1 TO, a pair of maps of halftone image fragments distribution Map PO – Map1 PO, a pair of maps of graphics fragments distribution Map GO – Map1 GO, and a pair of maps of background fragments distribution Map BO – Map1 BO.

On the next stage, with the help of logical XOR operation, difference between maps of background fragments distribution is calculated. Normally the differences are localized close to the borders of the background area and caused by the fact that a certain fragment contains some meaningful image data besides the background. For that sake, these fragments are classified again in a sliding window. The size of the window doesn't change. Sliding steps along X-axis and Y-axis are 1/8 of the window size in the corresponding dimension. The upper left corner of the window is shifted to up and to the left by 1/2 of the window size from the upper left corner of the fragment, as shown in Figure 4.

This reclassification helps to localize background areas. Distribution maps of background fragments are recalculated with source image size taken into account. Similar procedure is repeated for three other pairs.

The final segmentation of areas of each type is achieved by applying logical AND to refined TO, PO, and GO maps and to the inverted refined map of the background:

$$TO = TO \& (\sim BO)$$
$$PO = PO \& (\sim BO)$$
$$GO = GO \& (\sim BO)$$

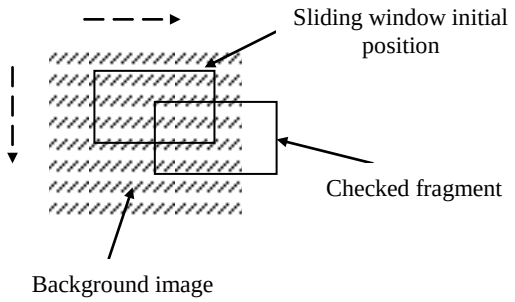


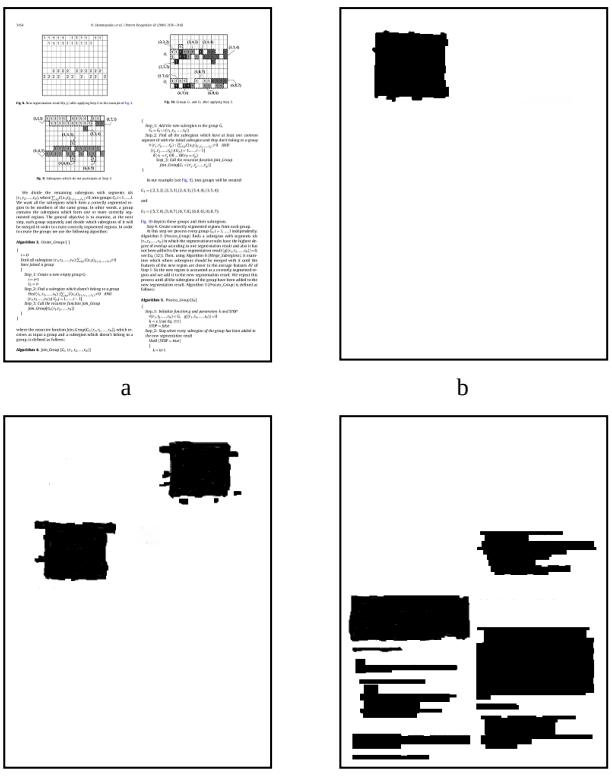
Figure 4: Starting position and moving direction of the sliding window

The last step of the segmentation is median filtering of the resulting binary maps. This processing yields maps of text areas and halftone images which can be used to extract images corresponding to the said areas.

3. EXPERIMENTAL RESULTS

A proposed method was implemented in C++ using OpenCV, an open image processing library.

The described algorithm can clearly be paralleled by data. After applying a wavelet transformation, each frequency matrix can be processed in parallel and independently from others. To that reason, parallel implementations of the algorithm were developed for the multicore CPU and for the computational cluster, utilizing standard means of parallel calculations such as OpenMP and MPI [6].



a b c d

Figure 5: A document segmentation. (a) Document image, (b) Map of graphics zone, (c) Map of half tone zone, (d) Map of text zone

The volume of the basic training set used to train the classifier equaled 1400 scanned documents. Among these were documents containing images of different types, as well as documents pertaining to one specific type only. The basic training set was enhanced by applying moderate brightness noise, rotating the image by angle of -5^0 to 5^0 and resizing the images of 10%.

Results of processing a document containing both text and graphics are presented in Figure 5.

Size of the test images vary from 870×1200 to 3400×4400, running time for 1800×2300 image is about 0.5 seconds on a 2HHz Pentium multicore processor for one thread implementation, and 0.16 seconds for 4 thread parallel implementation.

The proposed method consistently finds large contiguous sets of areas of the same type. However, sometimes it may fail to identify isolated text lines, erroneously classifying them as background. This drawback can be relaxed by applying a round of additional classification to the background area, using a window of smaller size.

4. ACKNOWLEDGEMENT

The author acknowledges Saint-Petersburg State University for a research grant 9.37.157.2014.

Research was carried out using computational resources provided by Resource Center "Computer Center of SPbU" (http://cc.spbu.ru).

REFERENCES

[1] Acharyya M., Kundu M.K. "Document image segmentation using wavelet scale-space features", Circuits and Systems for Video Technology, IEEE Transactions on (Volume:12 , Issue: 12) , 2002, pp. 1117 – 1127.

[2] Zhang Jing, Dong Wei, Zhang Youhui. "An Algorithm for Scanned Document Image Segmentation Based on Voronoi Diagram", Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on (Volume:1), 2012, pp. 156 – 159.

[3] H. S. Baird, M. A. Moll, Chang An, "Document Image Content Inventories", Proc. of SPIE/IS&T Document Recognition & Retrieval, 2007.

[4] F. Cesarini, S. Marinai, G. Soda, M. Gori. "Structured Document Segmentation and Representation by the Modified X-Y tree", International Conference on Document Analysis and Recognition, 1999, pp. 563.

[5] Cristianini Nello and Shawe Taylor, John. An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, 2000

[6] M. Firuziaan, O. Nommensen "Parallel Processing via MPI & OpenMP", Linux Enterprise, 10/2002