1994 International Conference on Pattern Recognition

Page Segmentation Using Decision Integration and Wavelet Packets

Kamran Etemad[†‡], David Doermann[†] and Rama Chellappa[†‡]
[†]Document Processing Group, Center for Automation Research
[‡]Department of Electrical Engineering
University of Maryland
College Park, MD 20742

Abstract

A new algorithm for layout-independent document page segmentation is suggested. Text, image and graphics regions in a document image are treated as three different "texture" classes. Soft local decisions on small blocks are made using wavelet packet based feature vectors. Segmentation is performed by propagating and integrating soft local decisions over neighboring blocks, within and across scales.The "uncertainties" associated with local decisions are reduced as more contextual evidence is incorporated in the process of decision integration. The majority, taken over weighted combined votes, determines the final decision. The suggested algorithm is based on parallel independent computations which have low complexity. It can also be applied to other signal and image segmentation tasks.

1 Introduction

Recent advances in information and communications technologies have increased the need for, and therefore the interest in, automated reading and processing of documents. Efficient storage and transmission of documents as well as archiving and information retrieval for document databases and "digital libraries" have become important research issues. For coding or understanding document images it is essential to identify text, image and graphics regions, as physical segments of the page, in order to be able to process them properly. Therefore the task of document page segmentation is one of the primary stages of most document processing systems. In this paper we describe a new method of layout independent physical page segmentation, in which few assumptions are made about the document's textual and graphical attributes or layout structure. The system is designed in such a way that, as hypotheses about document components are made and verified, more domain specific post-processing can occur.

Other methods of page segmentation and layout analysis described in the literature can be broadly classified into bottom-up and top-down approaches. Bottom-up techniques such as connected components [1] start from the pixel level and merge regions into larger and larger components (e.g. characters, words, text lines, paragraphs, etc.). In top-down methods the page is first split into blocks, and these blocks are identified and subdivided appropriately. For example, one might first locate columns and then split them into paragraphs, text lines, or even words. Examples of top-down methods include recursive projection profile cuts [2,3], run length smoothing or constrained run length [4] algorithms. There are also hybrid methods that combine the top-down and bottom-up approaches. In some approaches, after detecting major blocks, simple statistical tests classify them as text or non-text regions [3,5]. Black pixel density, black/white ratio or transitions, average vertical or horizontal run length, and cross-line correlations [5] are some of the features used in these post-classification stages.

Most of these techniques rely on prior knowledge or assumptions about generic document layout structure and textual and graphical attributes. Utilizing such knowledge results in simple, elegant and efficient page decomposition systems but also limits the range of applicability of the algorithm. Noise and degradations, multi-directional and curved text lines, touching or overlapping components, gray level or inverted text, complex layout structures, and differences in language, font size and other textual attributes are among

The support of this research by the Advanced Research Projects Agency (ARPA Order No. A550), under contract MDA 9049-3C-7217, is greatfully acknowledged.

the problems that limit most model-based approaches. Therefore in some applications it is desirable to have segmentation methods that do not rely heavily on a priori knowledge about the content or attributes of the document. Based on these observations, a texture segmentation based method for extracting text regions has been suggested [6]. The approach uses multi-channel Gabor filters as input features to a text region classifier.

In this paper, text, image and graphics regions in a document image are regarded as three classes of textures. A multi-scale wavelet packet representation, neural network based "soft classification", and decision integration are utilized to identify these regions on a page. A distinctive feature of this task, compared to other texture segmentation problems, is that, there are large intra-class as well as inter-class variations in the textural features. The multi-scale nature of document constituents (e.g. characters, lines) justifies the use of multi-scale representation and classification scheme.

The idea of soft decision integration is based on the fuzziness or uncertainty associated with local decisions about small windows. The uncertainties result from the limited view of the signals and/or the randomness and ambiguity inherent in the problem. For example, documents image blocks may contain text, image and graphics sub-regions that are adjacent to or overlap one another. In such situations it is not appropriate, even if we use an optimally designed classifier, to make hard (binary) local decisions. The inherent fuzziness of class memberships is not due to the noise or randomness; This supports our claim that soft local decisions yield more realistic and accurate membership values. The uncertainties reflected in the soft decisions are reduced by propagating and integrating decisions, made independently in the neighborhoods, within and across scales.

2 Wavelet Packet Basis

Wavelet packet analysis algorithms (wavelet transforms being special cases) allow us to perform adaptive Fourier windowing directly in the time domain by successive filtering of the signal into different frequency regions. The waveforms in the library are mutually orthogonal and each of them is orthogonal to all of its integer translates and dyadic rescaled versions. The full collection of wavelet packets (including translates and rescaled versions) provides us with a library of "templates" or "notes" which are matched "efficiently" to signals for analysis and synthesis.

We start with an exact Quadrature Mirror Filter (QMF) $h(n)$ satisfying

$$\sum h(n-2k)h(n-2\ell) = \delta_{k,\ell} \ , \ \sum h(n) = \sqrt{2}. \quad (1)$$

We let $g_k = h_{k+1}(-1)^k$ and define the operations F_i on $\ell^2(\mathbf{Z})$ into "$\ell^2(2\mathbf{Z})$" by

$$\begin{aligned} F_0\{s_k\}(i) &= 2\sum s_k h_{k-2i} \\ F_1\{s_k\}(i) &= 2\sum s_k g_{k-2i} \end{aligned} \quad (2)$$

The map $F(s_k) = F_0(s_k) \oplus F_1(s_k) \in \ell^2(2\mathbf{Z}) \oplus \ell^2(2\mathbf{Z})$ results in an orthogonal "decomposition", and satisfies the alias cancellation and perfect reconstruction conditions [7]. We now define the following sequence of functions.

$$\begin{cases} W_{2n}(x) = \sqrt{2}\sum h_k W_n(2x-k) \\ W_{2n+1}(x) = \sqrt{2}\sum g_k W_n(2x-k) \end{cases} \quad (3)$$

A *Wavelet Packet Basis* [7] for $L^2(\mathbf{R})$ is any orthonormal basis selected from the functions $2^{k/2}W_n(2^k t - j)$. The waveforms $\{W_\gamma\}$ are induced by three parameters $\gamma = \{k, n, j\}$ with physical interpretations of scale, frequency (or sequency), and position. These waveforms constitute our tree-structured basis. Each node in the tree represents a subspace of its parent's space and each subspace is the orthogonal direct sum of its two children. In wavelet analysis of 2-D signals, for simplicity, $L^2(R^2)$ is typically considered as a separable Hilbert space. In filter bank implementation of wavelet packets the assumption about separability of the signal space results in separable filters along the row and column directions [7]. Depending on the application, the choice of a suitable wavelet packet tree can be based on different criteria. Energy and entropy based [7] as well as class separability [8,9] based algorithms for basis selection have been suggested. In the following experiments, without any claim of optimality, a pyramidal wavelet transform and an energy based wavelet packet tree are used. In the following classification experiments, the wavelet-based features that are used are the second and third central moments (μ_2 and μ_3) of the image subbands computed over small windows W on each subimage of the transformed image.

3 Propagation and Integration of Local Soft Decisions

Many signal/image processing tasks consist of local processing of data followed by the combination of re-

sults obtained from the local windows. Image segmentation and boundary detection are examples of such tasks. On the other hand the recent success of neural networks and fuzzy systems has reconfirmed the fact that soft distributed decisions provide powerful and robust tools for dealing with uncertainty, ambiguity, and randomness [10,11].

Local decisions, which are inevitable in many signal processing schemes, suffer from such ambiguities and are not reliable on their own. There is therefore a need to devise methods of resolving the ambiguity and fuzziness of local decisions, in a consistent way, by incorporating more and more "evidence" or "contextual" information from the input signal or image. Since a reliable hard decision cannot be achieved by observation over a single small window, we suggest that local soft decisions/scores be propagated and integrated based on a weighting pattern in a neighborhood within each scale as well as across scales, and that the majority of the combined votes be used as a class identifier.

Assume that L classes of regions may be present in the image. For example, in the document domain we can set $L = 3$, corresponding to text, graphics and image regions. Our classifier is a map $F(.)$, typically non-linear and many-to-one, from the feature space $\{Y_s, s = (i,j)\}$ to the points in the "fuzzy" cube $[0,1]^L$. Thus

$$\begin{aligned} F : \Re^n &\rightarrow [0,1]^L \\ F(X_s) &= Y_s \end{aligned} \tag{4}$$

where Y_s is the vote vector whose i^{th} element is the score or fit value associated with class i. This framework, along with proper training, provides the soft local decisions.

We also assume that the information contained in a block B about a region A is proportional to the intersection area of B and A. As the image is analyzed using windows of size W and window shift steps of size w, the area contained in each block B is also partially covered by neighboring blocks. Since analysis is performed at several scales, the area in B is also covered by windows at each scale [11]. Therefore, the pattern of decision flow in space as well as in scale has to be determined.

Based on our assumptions, we define a "Vote Propagation Matrix" (VPM) M which shows how much the vote of the window centered at $s = (x,y)$ should affect or contribute to its neighbors:

$$M_{i,j} = 1-(|i|+|j|)/\delta+|ij|/\delta^2 \text{ for } -\delta < i,j < \delta \tag{5}$$

$$(V)_{x+i,y+j} = (V)_{x+i,y+j} + Y_{x,y} \cdot M_{i,j} \tag{6}$$

where $\delta = w/W$. The incremental vote $\Delta V_{x+i,y+j} = Y_{x,y} \cdot M_{i,j}$ is added to the vote of the corresponding principal block (centered at location (i,j) relative to the current block) to give its contribution to the sum of the votes in that block. After one complete scan of the image, the contributions of all neighboring blocks in support of matrix M have thus been accumulated, and matrix V_{tot} contains the combined vote for each block of size w. Note that the vote $(V)_{i,j}$ is a vector of L votes each corresponding to one class. All "decisions" thus far have been expressed as real numbers and no hard decision has been made.

To perform multi-resolution analysis we process the image at different scales and combine the resulting classifications. This is typically done from coarse to fine: we start at low resolution and use higher resolution where the confidence level is not satisfactory. The combination of decisions can make use of the fact that the windows of the low resolution image are projections of larger areas of the high resolution image. One can thus similarly define an "Across-Scales Vote Propagation Matrix"(ASVPM) based on overlapping areas and add the relevant weighted votes. The final majority votes and their confidence measures are based on the accumulation of all soft decisions within and across scales and the closeness of the best class candidates respectively:

$$(V_{final})_{i,j} = V^{(l*)} = max_l(V^{(l)})_{i,j} \tag{7}$$

$$Conf_{i,j} = V^{(l*)} - max_{l \neq l*}(V^{(l)})_{i,j} \tag{8}$$

4 Knowledge-Based Biased Voting

In some applications we may wish to incorporate constraints into our decisions based on a priori or derived knowledge about the domain. Although such constraints are often considered at higher levels of processing, one may also utilize them at the early stages of classification to get more reliable results. In the context of the described majority vote method this idea can easily be incorporated into the system, without increasing its complexity, by using a biased voting scheme. An expectation of observing a certain class of patterns in part of the scene is reflected in a biased vote in favor of a particular class within that region. In this case the system does not start from an all-zero vote matrix V_{tot}, rather, at each position, small non-zero initial votes are given to classes that have been frequently observed in that location. Thus we

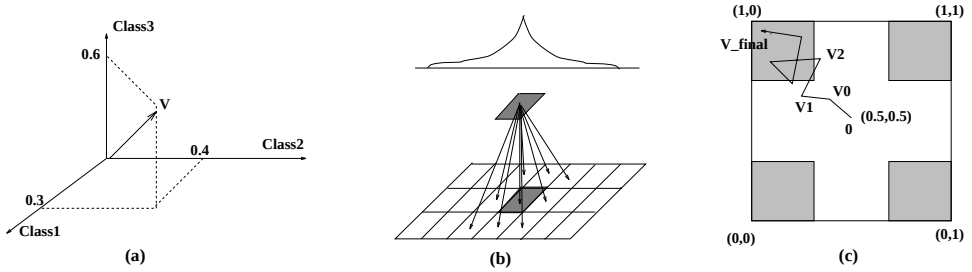


Figure 1: Decision integration: (a) Soft decision vectors, (b) Decision propagation profile in a neighborhood, (c) Weighted combination of soft decisions in the fuzzy decision square. Decisions outside the gray areas have low levels of confidence.

start from a non-zero vote matrix, on top of which the votes are accumulated. The initial vote may change during decision integration if the combined decision strongly favors another class. The method of "biased voting" based on prior expectations, can be viewed as not starting from the middle of the fuzzy decision cube (i.e. the most ambiguous point), but deviating from it toward one of the corners; see Figure 1c. Note that if a hard decision is made in each block, instead of a soft decision, the suggested combination of votes at each scale would be equivalent to a special form of median filtering on the vote matrix.

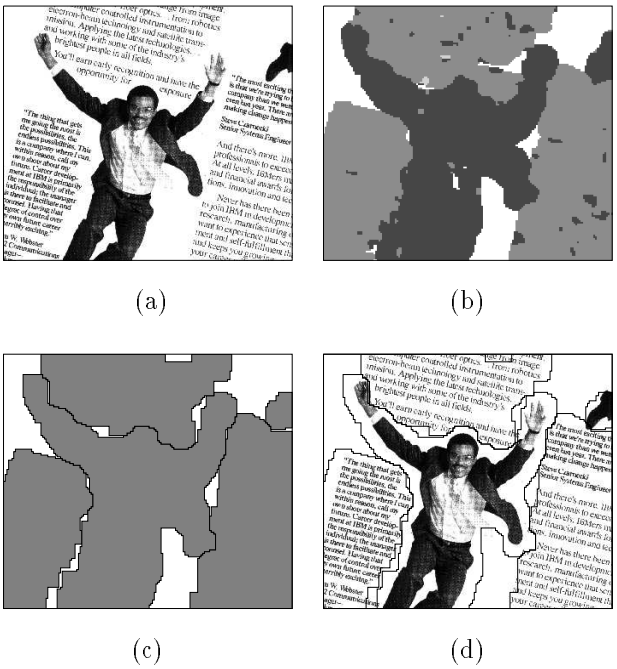
5 Postprocessing

The spatial patterns of combined soft decisions in the vote matrices directly reflect the locations, shapes and classes of major document blocks. In order to resolve ambiguity and obtain a more precise segmentation, additional postprocessing based on knowledge about the structure of the document may be required. The types of processing which can be applied fall into two classes - class dependent and class independent.

If the class of document (e.g. correspondence, journal article, advertisement, ...), can be identified, class dependent processing may proceed in a top-down manner using a hypothesize and test approach. We are able to use knowledge about the expected layout to refine the segmentation, as well as verify the document class hypothesis. As general case, suppose a two-column structure is inferred from a simple projection profile and a. Rectangular blocks are then be fit to the text blocks and precise column boundaries are identified. If an advertisement is hypothesized, polygonal or non-vertical rectangles may be used as an initial estimate. In our current system we apply a collection of rules which are derived from class independent observations about documents. By applying structural rules to the different regions, one can hypothesize and eliminate "logically undesirable" gaps and small noise-like mis-classified regions. We establish a set of rules which suggest that a text region is typically uniform, with no small inset graphic or image components. Graphic or image regions are "large" relative to the line spacing and/or font size. Graphic and image regions can have subordinate text, but such strings are assumed to be on the order of several symbols long. If text regions are touching graphic components, they are assumed to be part of the graphic component, and will be separated later. Similarly, graphic and image regions will not overlap and graphic regions having uniform (possibly white-space) backgrounds. These properties are enforced by applying morphological closing operations each region to filter out noise which is too small to constitute a valid document region. For text regions, a 10-12 step operation is sufficient to eliminate regions smaller then the largest symbols in a 9pt font, scanned at 300 dpi. For graphic and image regions, a larger kernel is used, eliminating slightly larger regions. The resulting regions are examined, and regions corresponding to "closed" holes are eliminated. Since each region is processed separately, there is no risk of merging multiple larger regions. When the segmentation component is incorporated into a more complete system which may include logical analysis, higher level processing, based on the local hypothesis, will also be used to identify rectangular structures, lines of text, and subordinate text regions when appropriate.

6 Experiments

To show the effectiveness of the soft decision integration method, we have applied it to document page segmentation. The input data consisted of several gray scale document pages scanned at 200 dpi.



(a) (b) (c) (d)

Figure 2: An example of segmentation results. (a) The original document image (b) Labeled regions using decision integration (c) Segmentation result after some postprocessing (d) The final segmentation.

The training set consisted of about 400 samples from each of the text, image and graphics sub-blocks. These samples were 16×16 pixel sub-blocks extracted randomly from several of the document pages. Ten wavelet channels that contained the highest a.c. energy were selected. A multi-layer feedforward neural network was used as the classifier [9]. The network consisted of 20 input, 8 hidden and three output units. The three outputs, corresponding to the classes text, image, and non-text, non-image, could take values in $[0, 1]$. In other words, any sub-block not identified as text or image was considered to be "graphics". Blank regions were detected separately in a straightforward way.

As mentioned before it is essential to train the classifier network in such a way that its outputs provide soft non-binary decisions about the class-memberships of the input image blocks. For any macro-pixel u and any class l one can compute the desired soft class-membership decision as

$$\forall u \in V \ (Targ)_l(x) = \frac{1}{|u|} \sum_{x \in u} I(Lab(x) = l) \qquad (9)$$

i.e., as the fraction of window area that has been labeled as l. Example results of soft decisions with and without the decision integration stage are shown in Figure 2. For post-processing, six steps of closing were applied to each component. This was sufficient to eliminate noise regions smaller than the area occupied by a capital "M" in a 9pt font.

7 Results and Discussion

The decision integration scheme has been successfully used to identify text, images, and graphics areas on a document page. Improved performance is obtained by making soft decisions. The decision integration method is advantageous when confident hard decisions cannot be made because of poor features, resolution or windowing problems, noise, or the inherent fuzziness of some classification tasks. The scheme has yielded good results using simple feature sets and classifiers. It is based on a approach that can be applied to other segmentation and classification problems. Almost all the computations and decisions are made independently and in parallel, without iterative stages. Thus the scheme is well adapted to fast implementation on distributed architectures.

8 References

[1] L.A. Fletcher and R. Kasturi,"A robust algorithm for for text string separation from mixed text/graphics images", IEEE Trans. PAMI, Vol. 10, pp. 910-918, 1988.

[2] D. Wang and S.N. Srihari,"Classification of newspaper image blocks using texture analysis", CVGIP, Vol. 47, pp. 327-352, 1989.

[3] M. Viswanathan and G. Nagy,"Characteristics of digitized images of technical articles", Proc. SPIE Vol. 1661, pp. 6-17, 1992.

[4] F.M. Wahl, K.Y. Wong and R.G. Casey,"Block segmentation and text extraction in mixed text image documents" CVGIP, Vol. 20, pp. 375-390, 1982.

[5] T. Pavlidis and J. Zhou,"Page segmentation and classification", CVGIP. Vol.54, pp. 484-496, 1992.

[6] A.K. Jain and S. Bhattacharjee,"Text segmentation using Gabor filers for automatic document processing", Machine Vision and Applications, Vol. 5, pp. 169-184, 1992.

[7] R.R. Coifman and M.V. Wickerhauser,"Entropy based algorithms for best basis selection", IEEE Trans. Information Theory, Vol. 38, pp. 713-718, 1992.

[8] K.Etemad and R. Chellappa ,"Separability based local basis selection for texture classification", Proc. ICIP, 1994.

[9] K. Fukunaga,"Introduction to Statistical Pattern Recognition", 2nd ed., Academic Press, 1990.

[10] B. Kosko, "Neural Networks and Fuzzy Systems", Prentice Hall, 1992.

[11] K.Etemad, R. Chellappa and D.Doermann,"Document page decomposition by integration of distributed soft decisions", Proc. ICNN, 1994.