DOCUMENT PAGE SEGMENTATION USING MULTISCALE CLUSTERING

Dipti Prasad Mukherjee and Scott T. Acton
Oklahoma Imaging Laboratory
School of Electrical and Computer Engineering
Oklahoma State University
Stillwater, OK 74078, USA
Email: sacton@okstate.edu

Abstract

The paper details a multiscale clustering technique for document page segmentation. In contrast to existing hierarchical (coarse-to-fine), multi-resolution methods, this image segmentation technique simultaneously uses information from different scaled representations of the original image. The final clustering of image segments is achieved through a fuzzy c-means based similarity measure between vectors in scale space. The segmentation process reduces the effects of insignificant detail and noise. Furthermore, object integrity is preserved in the segmentation process.

1. Introduction

Segmentation of scanned document pages into meaningful constituent regions is an effective precursor to document coding and content based retrieval. The objective is to identify each segmented region as a region representing printed text, graphics, halftones or continuous tone pictures, *etc.* Subsequent to image segmentation, knowledge based processing is needed to identify the constituent regions so that differential region specific document coding may be applied to individual block. In this paper, page segmentation is achieved using a multiscale clustering algorithm. Via clustering in the image *scale space*, multiscale features are extracted that cannot be obtained in any one particular scaled representation. In this context, we utilize a discrete scale space that includes a finite number of scaled representations for a given image, from fine to coarse.

For the problem of document segmentation, a series of scale space images is formed using morphological operators. Generation of the scale space is followed by a clustering process that simultaneously uses information from multiple scales. Typical classification processes that utilize only the original input image at a fixed feature scale will lead to intra-object classification errors (assigning two different class labels to the same object). The fixed scale methods are also sensitive to noise and insignificant detail. One of the important aspects of any thematic classification scheme is the selection of discriminating features. In this case, the scale space serves as a grid of feature vectors, allowing the detection of unique cohesive regions inside a scanned page image.

2. Background

Many earlier attempts to segment document images are relevant in light of the proposed approach. A survey of document segmentation and coding techniques is given in [6]. A notable approach for page classification is provided by Pavlidis and Zhou [7]. They have utilized the "white space" to segment a page image, followed by rule-based classification. Cross-correlation between streams of binary pixels along scan lines is used to discriminate between text, diagrams and halftones. In a recent article, de Queiroz and Eschbach [8] have segmented and classified JPEG-compressed documents. Morphological closing and opening operations are used to identify and crop individual segments in the compressed domain.

A number of multiscale classification and segmentation approaches are reported in the literature [2,3]. Here, multiscale representations are first computed. The segmentation process in these cases is strictly sequential, wherein *maximum a posteriori* (MAP) estimates at a given resolution are propagated from coarse to fine scales within a quadtree. Also, Bayesian segmentation techniques are used with the assumption that the distribution of pixel intensities is Markovian. Quadtree based multiscale segmentation [4] is also used to segment document pages. In this case, a trainable context model is required as ground truth for segmentation, and a class probability tree acts as the data structure for context model. In contrast to these coarse-to-fine methods, our approach attempts unsupervised multiscale classification without any *a priori* knowledge. A fuzzy *c*-means approach is used to group the scale space vectors for image segmentation.

3. Generation of scale space

The proposed approach has the following two major components:

0-7803-5467-2/99/ $ 10.00 © 1999 IEEE

- Formation of a scale space using a multiscale morphological operation.

- Application of an unsupervised fuzzy *c*-means clustering within the scale space: The corresponding pixels of each layer of the scale space act as elements in the feature vectors used in clustering.

Two morphological operations are investigated in this paper to derive multiple scales suitable for document segmentation. These are morphological opening and *area-based* morphological operators. The first approach generates a scale space using the basic morphological opening operation. In the second approach, area-based operators that remove features smaller than a specified area, regardless of shape are used to compute scaled images for document page segmentation.

Scale space properties of the multiscale morphological dilation-erosion are detailed in [5]. In our first approach, sequential combination of erosion followed by dilation is used to derive multiple scales. For a convex structuring element g_s, the morphological opening of image I at position x is given by:

$$(I \circ g_s)(x) = ((I \ominus g_s) \oplus g_s)(x). \quad (1)$$

For *n*th scale space representation, the structuring element g_s is the *n*-fold self-dilation of the structuring element g:

$$g_s|_n = g \oplus g \oplus g \cdots \oplus g. \quad (2)$$

We utilize the standard morphological operators such as (1) where the standard opening removes any positive-going image feature that is eclipsed by the structuring element (a square or circle, for example).

The rationale for investigating the standard morphological opening to construct the scale space is due to the relative computational efficiency with respect to other scale generating operators. At the same time, successive opening results in the desired merging of closely spaced text regions, while almost maintaining area of the gray tone image or graphics in white background.

Area open-close and area close-open are also used to generate scaled image representations for segmentation. Area operators have been applied previously to general filtering for image enhancement and image reconstruction [9]. The filters have the desirable property of connected invariance – connected components in the image level sets (that may correspond to image objects) are never partially removed in filtering. As a result, the filters may be used for noise and detail removal without boundary movement, and a straightforward method to scale an image is provided.

To apply area based operators to gray scale imagery, a stacking or threshold decomposition of the image is utilized. An image I can be decomposed into *level sets* I_t, where $I_t = \{x: I(x) \geq t\}$. The image intensity can be obtained from the level sets using

$$I(x) = \max\{t : x \in I_t\}. \quad (3)$$

This level set definition allows the area operators to be applied to each level set independently.

A level set is then defined as a binary image B with values of $B(x) \in \{0,1\}$ at any image location x. A connected component at x, $C_B(x)$ is defined as the set of locations y where there exists an unbroken path between x and y. Typically, 4-connectivity or 8-connectivity is used to define connected path in discrete domain images assuming rectangular tessellation. Then, the area-based operators remove these connected components depending on the scale of the object specified in terms of the area of the connected components.

For a set B, we can define the area open operation by

$$x \in \overset{a}{\circ}(B) \text{ if } |C_B(x)| \geq a, \quad (4)$$

where $|C_B(x)|$ is the cardinality (area in the discrete sense) of the connected component, and a is the minimum area. This implies that

$$x \notin \overset{a}{\circ}(B) \text{ if } |C_B(x)| < a. \quad (5)$$

Let the complement of the set B be denoted by B^c, where $B(x) = 0$, $B^c(x) = 1$. To form the area close operation, we have

$$x \in \overset{a}{\bullet}(B) \text{ if } |C_B^c(x)| \geq a, \quad (6)$$

where a denotes the minimum area of connected components in B^c. Of course,

$$x \notin \overset{a}{\bullet}(B) \text{ if } |C_B^c(x)| < a, \quad (7)$$

Combination of the area open and area close operators yields important scale-generating operators. The area open-close operation is given by $\overset{a}{\bullet}(\overset{a}{\circ}(B))$, while the area close-open operation is $\overset{a}{\circ}(\overset{a}{\bullet}(B))$. Both operations remove connected components of area less than a in both B and B^c, the complement of B. Thus, area open-close (AOC) and area close-open (ACO) can be

employed to control the minimum scale of objects in B. However, $\bullet_a(\circ_a(B)) \neq \circ_a(\bullet_a(B))$ in general. Once the area operators are applied independently at each level set, the image is reconstructed after pointwise stacking of all the level sets according to (3).

In constructing a viable scale space, the desirable properties include fidelity, Euclidean invariance, causality and strong causality. Each of these properties has important consequences for document page segmentation. For scale spaces created via area open-close and close-open, these properties are asserted. Fidelity ensures that segmentation of the scale space never introduces a segment that is not originally present in the input image. Euclidean invariance allows generation of the identical scale space irrespective of translation and rotation of the input image. Causality constraints the dependence of image at any scale to its immediately previous scale. Strong causality prevents drifting of edges within fine to coarse representation of the image.

3.1 Scale space classification

Within a scale space $\{\mathbf{I}\}$, we inspect the intensity $I_s(x)$ at position (x) and scale s through scale space. We call $\mathbf{I}(x)$ the scale space vector at (x); it is one-dimensional signal parameterized by scale s. The scale space classifier clusters pixels in the input imagery using the similarity between scale space vectors.

Within the fuzzy c-means (FCM) clustering algorithm [1], the familiar least-squared error criterion is applied:

$$J_m(U,\mu) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m \| d_{ik} \|. \qquad (8)$$

Here, U is the fuzzy c partition of the scale space, and m is the fuzzy exponent where $m \in [1,\infty]$. For the n scale space vectors $\{\mathbf{I}\}$, the measure $\| d_{ik} \| = \| \mathbf{I}_k(x) - \mu_i \|$ is the distance between the ith cluster center μ_i and kth scale space vector $\mathbf{I}_k(x)$. In the fixed scale case, this distance is computed between the intensity of the original image $I(x)$ and the cluster center μ_i. This distance is weighted by the fuzzy membership value u_{ik} corresponding to ith class and kth scale space vector. The error criterion $J_m(U,\mu)$ is minimized subject to the conditions $\sum_{i=1}^{c} u_{ik} = 1$, $0 < \sum_{k} u_{ik} < k$ and $u_{ik} \geq 0$. The fuzzy membership value is updated according to

$$u_{ik} = 1/[\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{2/(m-1)}]. \qquad (9)$$

The cluster center is updated for all the classes according to

$$\mu_i = \frac{\sum_{k=1}^{n}(u_{ik})^m \mathbf{I}_k}{\sum_{k=1}^{n}(u_{ik})^m}. \qquad (10)$$

The algorithm is terminated when insignificant changes between two consecutive μ_i values are observed.

4. Results

The proposed algorithm is tested on a number of scanned document pages. An example is shown in Fig. 1. The input image contains noise and artifacts due to the scanning process, including the lighter text that appears from the backside of the scanned page. The result of applying fixed scale FCM clustering on Fig. 1 is shown in Fig. 2. The classes in this case consist of the foreground and background. The result in Fig. 2 emphasizes insignificant details, and regions corresponding to printed text and graphics are erroneously labeled as part of background.

A salt and pepper noise of density in the order of 0.02 is added to the original image of Fig. 1. The resultant image is shown in Fig. 3. The result of AOC operation on Fig. 3 is shown in Fig. 4. The scale determined by the length of the connected component is fixed at 5. As expected, the AOC operation has effectively removed noises and undesired features maintaining the fidelity of the text, graphics and picture blocks of the original image. At the same time, the contrast between text region and the inter-line white spaces within the text block is reduced so that the whole block could be cropped as a text block.

In contrast to Fig. 2, the scale space classification result in Fig. 5 gives a meaningful separation between foreground and background. In this case, the scale space is constructed via successive opening of Fig. 4 using square structuring elements of increasing size (1x1, 2x2, 3x3, and 4x4). The multiscale clustering approach reduces classification error and provides cohesive regions for document coding.

Fig. 6 shows another document page with watermark. The fixed scale 2-class FCM classification result is shown Fig. 7. This approach clearly fails in extracting the text as cohesive regions. Further degraded version of Fig. 6 with salt and pepper noise is shown in Fig. 8. The AOC filtered image for scale/area of 5 is shown in Fig. 9. Note that the area open close operation combines the watermark and text body. Fig. 10 shows the result of scale space classification where scale space is generated through successive opening operation with increasing structuring element size (1x1, 2x2 and 3x3). Segments obtained in Fig. 10 are better suited for document coding, as compared to the results of Fig. 7.

For further comparison, the classification results are compared with the ground truth data and the classification accuracy is calculated both for fixed scale and scale space classifier. The comparison is shown in Table I. Clearly the scale space classifier outperforms the result obtained by fixed scale classifier.

In summary, the scale space allows meaningful page segmentation effective for document coding. The drawback, however, is the additional cost incurred in generating the scale space itself. We are presently investigating effective data structures to efficiently compute the scale space. We are also pursuing object based coding algorithms that complement the segmentation approach.

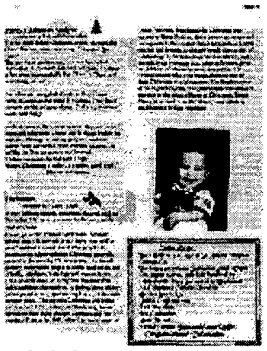

Fig. 1: Original document image.

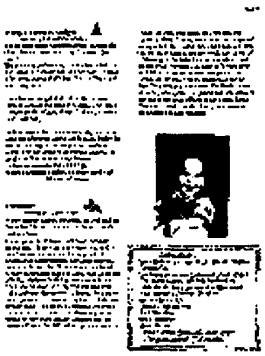

Fig. 2: 2-class FCM of Fig. 1.

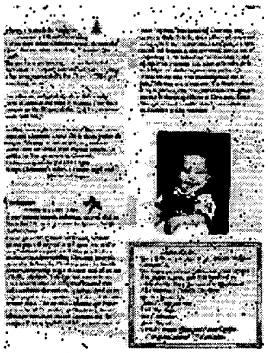

Fig. 3: Original image (Fig. 1) with salt and pepper noise (noise density=0.02).

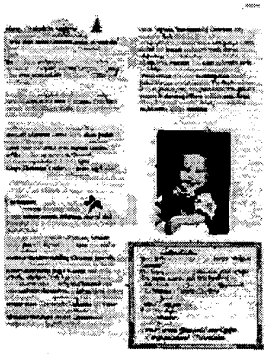

Fig. 4: After AOC operation on Fig. 3 at scale 5.

Fig. 5: FCM clustering (2 classes) using morphological scale space.

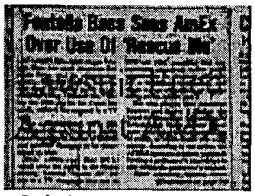

Fig. 6: Original document image.

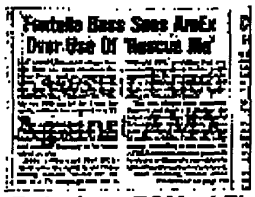

Fig. 7: 2-class FCM of Fig. 6.

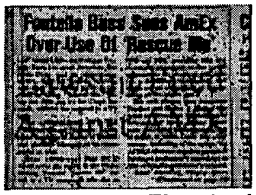

Fig. 8: Original image (Fig. 6) with salt and pepper noise (noise density=0.02).

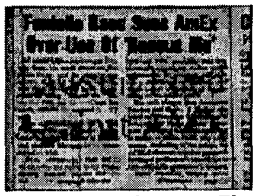

Fig. 9: After AOC operation on Fig. 8 at scale 5.

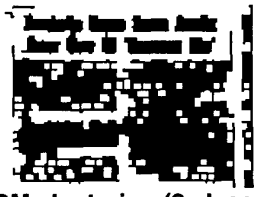

Fig. 8: FCM clustering (2 classes) using morphological scale space.

Table I: Comparisons of classification accuracy (FSFCM: fixed scale fuzzy *c*-means, SSFCM: scale space fuzzy *c*-means)

	FSFCM (%)	SSFCM (%)
Fig. 1 Example	61.3	88.2
Fig. 6 Example	61.9	80.4

REFERENCES

[1] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.

[2] C. Bouman and B. Liu, "Multiple Resolution Segmentation of Textured Images," *IEEE Trans. PAMI*, vol. 13, no. 2, pp. 99-113, 1991.

[3] C.A. Bouman and M. Shapiro, "A Multiple Random Field Model for Bayesian Image Segmentation," *IEEE Trans. Image Processing*, vol. 3, no. 2, pp.162-177, 1994.

[4] H. Cheng and C.A. Bouman, "Trainable Context Model for Multiscale Segmentation," *Proceedings Int. Conf. on Image Processing*, Chicago, October, 1998.

[5] P.T. Jackway and M. Deriche, "Scale Space Properties of the Multiscale Morphological Dilation-Erosion," *IEEE Trans. PAMI*, vol. 18, no. 1, 1996.

[6] M. Nadler, "A Survey of Document Segmentation and Coding Techniques," *CVGIP*, vol. 28, pp. 240-262, 1984.

[7] T. Pavlidis and J. Zhou, "Page Segmentation and Classification," *CVGIP: Graphical Models and Image Processing*, vol. 54, pp. 484-496, 1992.

[8] R.L. de Queiroz and R. Eschbach, "Fast Segmentation of the JPEG Compressed Documents," *Journal of Electronic Imaging*, vol. 7, no. 2, pp. 367-377, 1998.

[9] P. Salembier and J. Serra, "Flat Zones Filtering , Connected Operators, and Filters by Reconstruction," *IEEE Trans. Image Processing*, vol. 4, no. 8, pp. 1153-1160, August 1995.