Scanned Color Document Image Segmentation Using the EM Algorithm

John C. Handley
Xerox Corporation
800 Phillips Road, MS 128-27E
Webster, NY 14580 USA
Jhandley@crt.xerox.com

Abstract

A robust, efficient scanned color document segmentation algorithm is presented that performs a three-dimensional (3D) thresholding of color pixels. At the heart of the algorithm is the Expectation-Maximization (EM) algorithm which fits a mixture of two 3D gaussians to L* a* b* color data sampled from pixels in the image to separate foreground and background. The thresholding process uses a quadratic boundary to produce a selector plane that indicates, for each pixel, whether it belongs to foreground or background. The binary selector plane is further processed and analyzed to extract objects such as photographs or graphics. Segmented document images are then encoded in a mixed raster content format for efficient compression.

Introduction

Although much content is online, there remains a substantial amount of information in paper documents. Workflows can require extracting information in printed forms, converting legacy documents, or committing content of paper documents to a storage and retrieval system. In document processing systems, scanning completes the cycle: electronic, print, electronic.

Conversion of printed documents to electronic format has been the subject of thousands of research articles and numerous books [1]. Most work has focused on binary black and white documents. Yet the majority of documents today are in color at increasingly higher resolutions.

From a document interchange perspective, the mixed raster content (MRC) imaging model enables representation of basic document structure [2]. Its intent is to facilitate high compression by segmenting a document image into a number of regions according to compression type. For example, text pixels are extracted and encoded with ITU-T G4 or JBIG2. Background and pictures are extracted and compressed with JPEG (perhaps at differing quantization levels). Thus a document image is partitioned into a number of regions according to appropriate compression schemes. But MRC can also describe a basic "functional" decomposition of the image: text, background, photographs, and graphics, which can be used for subsequent processing. For example, text can be "OCRed" or photographs color corrected for different display media.

We describe a color document image segmentation system that has the following advantages: 1) it is robust and adaptive to a multitude of scanners and "show through" by virtue of the unsupervised "clustering" algorithm at its core; 2) it is simple and fast, making it amenable to software execution; and 3) it reduces much of the document analysis problem to processing binary images.

The heart of the segmentation system is an expectation-maximization algorithm to fit a mixture of three-dimensional gaussians to L*a*b* pixel samples. From the estimated densities and proportionality parameter, a quadratic decision boundary is calculated and applied to every pixel in the image. A binary selector plane is maintained that assigns one to the selector pixel value if the pixel is foreground and zero otherwise (background). The component distribution with the greater luminance is assigned the role of a background prototype. This process is essentially 3D thresholding. If the Euclidean distance of the estimated means are close together or if the estimated proportionality parameter is near zero or one, the samples fail to exhibit a clear mixture - the sample is homogenous or is not well-fitted with a mixture of 3D gaussians. At this stage, a segmentation attempt is made using only the L* channel by a mixture of 1D gaussians. Again, if estimated means are close or estimated proportionality parameter close to zero or one, the segmenter reports that the document image cannot be segmented.

Next, the selector is processed to find connected components by first doing a morphological opening and then a closing. Large connected components are extracted as objects and output as foreground/mask pairs. The segmented document image is now ready for subsequent processing. The objects may be smoothed or enhanced according to image type, the selector plane subjected to further analysis

as a binary document image, etc. Also, one may compress the image according to the TIFF-FX profile M standard or variant.

EM in a Nutshell

Expectation-Maximization (EM) is a general technique for maximum-likelihood estimation when data are missing. The seminal paper is Dempster, Laird, and Rubin [3] and a recent comprehensive treatment is McLachlan and Krishnan [4]. The mixture-of-gaussians (MoG) estimation problem is a straightforward and intuitive application of EM.

There are other approaches to this problem. Estimating the MoG can be thought of as unsupervised pattern recognition.

Consider two multivariate normal distributions $f_i(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$, $i=1,2$. The MoG distribution is

$$f(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2)=\alpha f(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)+(1-\alpha)f(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$$

where $0\le\alpha\le 1$ is the proportionality parameter. Given an i.i.d sample $\mathbf{x}=\{\mathbf{x}_i;i=1,\ldots,N\}$ from *f*, one would like to compute maximum likelihood estimates of the proportion, the vector means and covariance matrices. Unfortunately, no closed form is known (unlike the homogeneous case). One must maximize the likelihood numerically,

$$L(\mathbf{x};\alpha,\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)=\prod_{i=1}^{N}[\alpha f(\mathbf{x}_i;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)+(1-\alpha)f(\mathbf{x}_i;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)] \tag{1}$$

The EM algorithm provides an iterative and intuitive method to produce mles.

The missing data in this case is membership information. Let $Z_{ij}=1$ if $\mathbf{x}_j$ is from $f(\cdot;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$, and zero otherwise, $i=1,2$. The unobserved random variable Z_{ij} indicates to which distribution the observation belongs: $P(Z_{1j}=1)=\alpha$. Were, in fact, Z_{ij} observed, we could form mles. Let $Z_{ij}=z_{ij}$ and form the likelihood

$$L(\mathbf{x};\alpha,\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)=\prod_{j=1}^{N}\left[\alpha f(\mathbf{x}_j;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)\right]^{z_{1j}}\times\left[(1-\alpha)f(\mathbf{x}_j;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)\right]^{z_{2j}} \tag{2}$$

which yields mles

$$\hat{\alpha}=\frac{1}{N}\sum_{j=1}^{N}z_{1j} \tag{3}$$

$$\hat{\boldsymbol{\mu}}_i=\frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_{ij}\Big/\sum_{j=1}^{N}z_{ij},\ i=1,2 \tag{4}$$

and covariance mles omitted for brevity.

If we new the parameter values, we could estimate z_{ij} by conditional expectations

$$\hat{z}_{ij}=E(Z_{ij}\mid\alpha,\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)=\frac{f(x_j;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)}{\alpha f(x_j;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)+(1-\alpha)f(x_j;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)} \tag{5}$$

The first step in the EM algorithm is to initialize parameter estimates, $\hat{\alpha}^{(0)},\hat{\boldsymbol{\mu}}_1^{(0)},\hat{\boldsymbol{\Sigma}}_1^{(0)},\hat{\boldsymbol{\mu}}_2^{(0)},\hat{\boldsymbol{\Sigma}}_2^{(0)}$. The next step, the "E-step," is to use Eq. 5 to get estimates of the z_{ij}. The next step, the "M-step" is to use these estimates of the z_{ij} and the original data in Eqs. (3) and (4) to get updated mles of the parameters. The algorithm iterates these two steps until some measure of convergence is achieved (typically, updated parameter estimates differ little from previous ones, or the likelihood value stabilizes). That's essentially all there is to it for MoG. The fact that such a simple and intuitive method works under general conditions is a major achievement in late 20th century statistics.

Mixed Raster Content

Document image segmentation is done for a number of reasons. Recently there has been interest in segmenting a document image for compression. In this case, segmentation classes are compression classes, i.e., regions amenable to compression with appropriate algorithms: text with ITU-T Group 4 (MMR) and color images with JPEG [5, 6, 7]. One advantage of this approach is the one avoids compressing text with JPEG where it is known to produce ringing and mosquito noise. One can also use segmentation to find rendering classes, e.g., halftone regions to be descreen text to be sharpened, and photos to be enhanced.

Mixed raster content is an imaging model directed toward facilitating compression, yet can be used as a "carrier" for documents segmented for rendering or layout analysis.

Formally, we represent a color image as a mapping from a raster to a triplet of 8-bit colors:

$$I:[m_x,n_x]\times[m_y,n_y]\rightarrow[0,255]^3$$

where $0 \le m_x < n_x$ and $0 \le m_y < n_y$. A 3-plane mixed raster content representation uses a mask M to separate background and foreground content. Let $m_x = m_y = 0$ and

$$M0 : [0, n_x] \times [0, n_y] \to \{0,1\}$$

be a binary mask where n_x and n_y represent the complete extent of the image raster. Let

$$FG0, BG0 : [0, n_x] \times [0, n_y] \to [0,255]^3$$

be foreground and background images, respectively. A 3-plane MRC document image representation is

$$I(x, y) = (1 - M0(x, y))BG0(x, y) + M0(x, y)FG0(x, y)$$

for $(x, y) \in [0, n_x] \times [0, n_y]$. Essentially, a (vector) pixel value is selected from the background, if the mask is zero, and from the foreground if the mask is one. One can view the imaging operation as pouring the foreground through a mask onto the background.

We also need the concept of an object, which is a foreground/mask pair meant to represent a photograph or graphic. An object foreground is an image *FGi* and a mask *Mi*:

$$FGi : [mi_x, ni_x] \times [mi_y, ni_y] \to [0,255]^3$$
$$Mi : [mi_x, ni_x] \times [mi_y, ni_y] \to \{0,1\}$$

where $0 \le mi_x < ni_x \le n_x$ and $0 \le mi_y < ni_y \le n_y$. An object is imaged by $O_i(x, y) = Mi(x, y)FGi(x, y)$ for $(x, y) \in [mi_x, ni_x] \times [mi_y, ni_y]$ and zero elsewhere. The number of objects that can appear on a page is not *a priori* restricted except that objects cannot overlap (for we cannot segment them if they do) and they must have a certain minimum area (say, 2 square inches). The final document raster is imaged as

$$I(x, y) = (1 - M0(x, y))BG0(x, y) + M0(x, y)FG0(x, y) + \sum_{i=1}^{N} O_i(x, y)$$

This decomposition is by no means unique and there are others more appropriate for compression [8, 9, 10].

Segmentation Algorithm

1) Read a raster image into memory
2) Convert it to L*a*b*
3) Sample the image at a number of points uniformly distributed points
4) Using the Expectation-Maximization (EM) algorithm to estimate a mixture parameter, two 3D means and covariance matrices: $\hat{\alpha}, \hat{\boldsymbol{\mu}}_f, \hat{\boldsymbol{\Sigma}}_f, \hat{\boldsymbol{\mu}}_b, \hat{\boldsymbol{\Sigma}}_b$ purportedly representing foreground and background gaussians; i.e., the data are fit with $\alpha f(\mathbf{x}; \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b) + (1-\alpha) f(\mathbf{x}; \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$, where $\mathbf{x} = (l^*, a^*, b^*)$ at a point.
5) If $\left\| \hat{\boldsymbol{\mu}}_b(l^*) - \hat{\boldsymbol{\mu}}_f(l^*) \right\| > t$ and $s_1 \le \hat{\alpha} \le s_2$ then foreground and background are well-separated in L*a*b*
    a. For each pixel $\mathbf{x}$ in the image, if $\hat{\alpha} f(\mathbf{x}; \hat{\boldsymbol{\mu}}_b, \hat{\boldsymbol{\Sigma}}_b) > (1-\hat{\alpha}) f(\mathbf{x}; \hat{\boldsymbol{\mu}}_b, \hat{\boldsymbol{\Sigma}}_b)$ put $\mathbf{x}$ in the background and put a 0 in the mask M0 at that point; else put $\mathbf{x}$ in the foreground and put a 1 in the mask M0 at that point.
    b. Make a copy S of the mask M0.
    c. Convert S to horizontal runlengths and do a closing with a horizontal element (this closes small gaps)
    d. Convert S to vertical runlengths and do a closing with a vertical element (this closes small gaps)
    e. Convert S to horizontal runlengths and do an opening with a horizontal element (this smoothes window boundaries)
    f. Convert S to vertical runlengths and do an opening with a vertical element (this smoothes window boundaries)
    g. Convert S to connected components.
    h. For each connected component *Mi* larger than *thresh* in area
        i. Remove *Mi* from *M*0
        ii. Mask out *Mi* from *FG*0 making *FG*0 white where *Mi* is 1 and copying those pixels to a new object foreground *FGi*
        iii. Fill the holes in *Mi* by
            1. Finding small connected components in M*i* of 0-valued pixels
            2. Painting those connected components 1.
        iv. Output the found object as a foreground/mask pair (*FGi*,*Mi*)

i. Output the background $BG0$, the mask (selector) $M0$, and foreground $FG0$

6) If $\left\|\hat{\boldsymbol{\mu}}_b(l^*) - \hat{\boldsymbol{\mu}}_f(l^*)\right\| \le t$ and $s_1 \le \hat{\alpha} \le s_2$ then fit a 1D mixture of gaussians to the L* values and perform step 5 (which can be reduced to a simple threshold operation) .

7) Else the data form one gaussian blob or the EM algorithm failed to return a reasonable estimate, return the original image as $BG0$.

Implemented in software with unoptimized code, the current version runs in about 60 seconds on a Sun Sparc Ultra 60.

Some Segmentation Results

Figure 1 shows an magazine page scanned on a UMAX scanner at 400 x 400 dpi (left) and the background extracted using the EM-based segmentation algorithm. There is a considerable amount of show-through that could confound a simple thresholding.

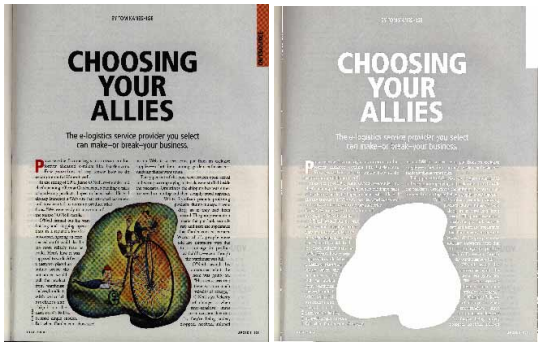

Figure 1. Scanned color document (left) and background, BG0 (right).

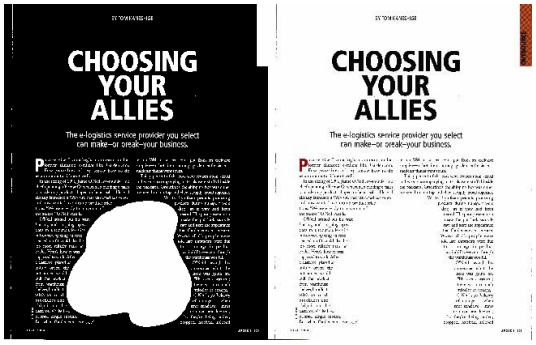

Figure 2. Mask M0 (left), foreground FG0 (right).

Figure 3. Object: mask $M1$ (left), foreground $FG1$ (right)

Summary

The EM algorithm provides a simple and robust procedure for color document image segmentation. It is particularly useful in producing an MRC representation with a background, a mask layer (containing text and line art), a foreground, and a number of foreground/mask pairs representing objects. It is robust to different backgrounds and scanner characteristics, including document images with a significant amount of show-through.

References

1. J. C. Handley, "Document Recognition," in *Electronic Imaging Technology*, E. R. Dougherty, ed., SPIE Press, Bellingham, 1999.
2. R. de Queiroz, R. Buckley, M. Xu, Mixed raster content (MRC) model for compound document image compression, *Proc. SPIE*, *Visual Communications and Image Processing*, vol. 3653, pp. 1106-1117, Jan 1999.
3. A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society B*, 39, pp. 1-38 (1977).
4. G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, Wiley, New York (1997).
5. L. Bottou, P. Haffner, P. G. Howard, P. Simard, Y. Bengio, Y. LeCun, High quality document image compression with "DjVu," *Journal of Electronic Imaging*, 7(3), pp. 410-425 (1998).
6. ITU-T Recommendation T.6, Facsimile coding schemes and coding control functions for group 4 fascimile apparatus, Nov. 1988.
7. ITU-T Recommendation T.44, Mixed Raster Content (MRC).
8. D. Mukherjee, N. Memin, and A. Said, JPEG-Matched MRC compression of compound documents, *ICIP 2001*, pp. 434-437.
9. H. Cheng and C. Bouman, Document compression using rate-distortion optimized segmentation, *Journal of Electronic Imaging*, 10(2), 2001, pp. 460-474.
10. R. de Queiroz, Z. Fan, T. D. Tran, Optimizing block-thresholding segmentation for multilayer compression of compound images, *IEEE Transactions on Image Processing*, 9(9), 2000, pp. 1461-1471.