ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

Test Segmentation of MRC Document Compression and Decompression by Using MATLAB

N.Rajeswari[1], S.Rathnapriya[2], S.Nijandan[3]

Assistant Professor/EEE, GRT Institute of Engineering & Technology, Tamilnadu, India[1]

Assistant Professor/ECE, GRT Institute of Engineering & Technology, Tamilnadu, India[2, 3]

Abstract- The mixed raster content (MRC) standard specifies a framework for document compression which can dramatically improve the compression/ quality tradeoff as compared to traditional lossy image compression algorithms. The key to MRC compression is the separation of the document into foreground and background layers, represented as a binary mask. Therefore, the resulting quality and compression ratio of a MRC document encoder is highly dependent upon the segmentation algorithm used to compute the binary mask. The incorporated multi scale framework is used in order to improve the segmentation accuracy of text with varying size. In this paper, we propose a novel multi scale segmentation scheme for MRC document encoding based on the sequential application of two algorithms. The first algorithm, cost optimized segmentation (COS), is a block wise segmentation algorithm formulated in a global cost optimization framework. The second algorithm, connected component classification (CCC), refines the initial segmentation by classifying feature vectors of connected components using a Markov random field (MRF) model. The combined COS/CCC segmentation algorithms are then incorporated into a multi scale framework in order to improve the segmentation accuracy of text with varying size.

Index Terms—MRC, COS, CCC, MRF.

I. INTRODUCTION

With recent advances in data processing systems and in electronic imaging and scanning devices, documents are now present in a wide spectrum of printing systems. From offset printers to home desktop computers, documents in digital form became common place. Frequently, documents are available as bitmaps and may contain text, graphics and pictures. As typical documents are often generated at a reasonably high resolution, document image sizes are invariably large and commonly consume several megabytes for storage. Furthermore, the final destinations for those documents are frequently parties other than those who generated them. Thus, it is desirable to possess the ability to transmit those large document images. Storage or transmission of large amounts of data is often costly and image compression is a necessity. Many standard compression algorithms are available today and in common use commercially. More are continually being developed to improve on existing methods or to meet special requirements. As a rule, any one compression algorithm was developed with a particular image type and characteristic, and a particular application in mind. For a different image type or application, a given algorithm either does not apply or does not perform as well as some other, better-tailored algorithm. No single algorithm is best across all image types or applications.

For example, a typical color document scanned at 300 dpi requires approximately 24M bytes of storage without compression. While JPEG and JPEG2000 are frequently used tools for natural image compression, they are not very effective for the compression of raster scanned compound documents which typically contain a combination of text, graphics, and natural images. This is because the use of a fixed DCT or wavelet transformation for all content typically results in severe ringing distortion near edges and line-art.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

The mixed raster content (MRC) standard is a framework for layer-based document compression defined in the ITU-T T.44 [1] that enables the preservation of text detail while reducing the bitrate of encoded raster documents. The most basic MRC approach, MRC mode 1, divides an image into three layers: a binary mask layer, foreground layer, and background layer. The binary mask indicates the assignment of each pixel to the foreground layer or the background layer by a "1" (Black) or "0" (White) value, respectively. Typically, text regions are classified as foreground while picture regions are classified as background. Each layer is then encoded independently using an appropriate encoder. For example, foreground and background layers may be encoded using traditional photographic compression such as JPEG or JPEG2000 while the binary mask layer may be encoded using symbol-matching based compression such as JBIG or JBIG2. Moreover, it is often the case that different compression ratios and sub sampling rates are used for foreground and background layers due to their different characteristics. Typically, the foreground layer is more aggressively compressed than the background layer because the foreground layer requires lower color and spatial resolution. Figure 1 shows an example of layers in an MRC mode 1 document.

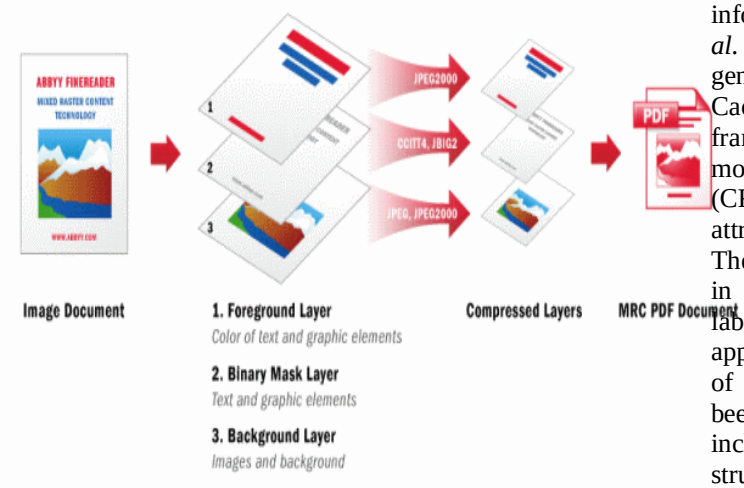

Fig 1 : Illustration of Mixed Raster Content (MRC) document compression standard mode 1 structure.

Perhaps the most critical step in MRC encoding is the segmentation step, which creates a binary mask that separates text and line-graphics from natural image and background regions in the document. Segmentation influences both the quality and bitrate of an MRC document. For example, if a text component is not properly detected by the binary mask layer, the text edges will be blurred by the background layer encoder. Alternatively, if non-text is erroneously detected as text, this error can also cause distortion through the introduction of false edge artifacts and the excessive smoothing of regions assigned to the foreground layer. Furthermore, erroneously detected text can also increase the bit rate required for symbol-based compression methods such as JBIG2. This is because erroneously detected and unstructured non-text symbols are not be efficiently represented by JBIG2 symbol dictionaries.

Many recent approaches to text segmentation have been based on statistical models. One of the best commercial text segmentation algorithms, which are incorporated in the DjVu document encoder, uses a hidden Markov model (HMM) [2], [3]. The DjVu software package is perhaps the most popular MRC-based commercial document encoder. Although there are other MRC-based encoders such as LuraDocument [4], we have found DjVu to be the most accurate and robust algorithm available for document compression. However, as a commercial package, the full details of the DjVu algorithm are not available. Zheng *et al.* [5] used an MRF model to exploit the contextual document information for noise removal. Similarly, Kumar [6] *et al.* used an MRF model to refine the initial segmentation generated by the wavelet analysis. J. G. Kuk *et al.* and Cao *et al.* also developed a MAP-MRF text segmentation framework which incorporates their proposed prior model [7], [8]. Recently, a conditional random field (CRF) model, originally proposed by Lafferty [9], has attracted interest as an improved model for segmentation. The CRF model differs from the traditional MRF models in that it directly models the posterior distribution of labels given observations. For this reason, in the CRF approach the interactions between labels are a function of both labels and observations. The CRF model has been applied to different types of labeling problems including block wise segmentation of manmade structures [10], natural image segmentation [11], and pixel wise text segmentation [12].

In this paper, we present a robust multiscale segmentation algorithm for both detecting and segmenting text in complex documents containing background gradations, varying text size, reversed contrast text, and noisy backgrounds. While considerable

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

research has been done in the area of text segmentation, our approach differs in that it integrates a stochastic model of text structure and context into a multiscale framework in order to best meet the requirements of MRC document compression. Accordingly, our method is designed to minimize false detections of unstructured non-text components (which can create artifacts and increase bit-rate) while accurately segmenting true-text components of varying size and with varying backgrounds. Using this approach, our algorithm can achieve higher decoded image quality at a lower bit-rate than generic algorithms for document segmentation. We note that a preliminary version of this approach, without the use of an MRF prior model, was presented in the conference paper of [13], and that the source code for the method described in this paper is publicly available. 1

Our segmentation method is composed of two algorithms that are applied in sequence: the cost optimized segmentation (COS) algorithm and the connected component classification (CCC) algorithm. The COS algorithm is a block wise segmentation algorithm based on cost optimization. The COS produces a binary image from a gray level or color document; however, the resulting binary image typically contains many false text detections. The CCC algorithm further processes the resulting binary image to improve the accuracy of the segmentation. It does this by detecting non-text components (i.e. false text detections) in a Bayesian framework which incorporates a Markov random field (MRF) model of the component labels. One important innovation of our method is in the design of the MRF prior model used in the CCC detection of text components. In particular, we design the energy terms in the MRF distribution so that they adapt to attributes of the neighboring components' relative locations and appearance. By doing this, the MRF can enforce stronger dependencies between components which are more likely to have come from related portions of the document.

II. COST OPTIMIZED SEGMENTATION (COS)

The Cost Optimized Segmentation (COS) algorithm is a block-based segmentation algorithm formulated as a global cost optimization problem. The COS algorithm is comprised of two components: block wise segmentation and global segmentation. The block wise segmentation divides the input image into overlapping blocks and produces an initial segmentation for each block. The global segmentation is then computed from the initial segmented blocks so as to minimize a global cost function, which is carefully designed to favor segmentations that capture text components. The parameters of the cost function are optimized in an off-line training procedure. A block diagram for COS is shown in Fig. 2.

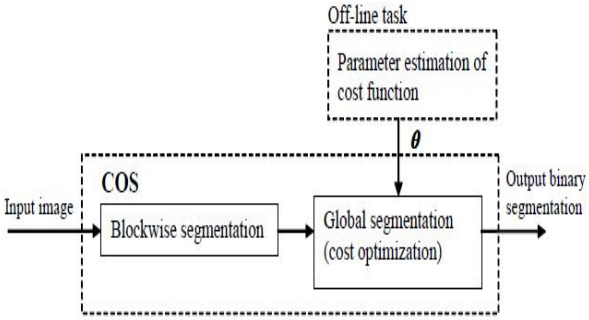

Fig. 2. The COS algorithm comprises two steps: block wise segmentation and global segmentation. The parameters of the cost function used in the global segmentation are optimized in an off-line training procedure.

A. Blockwise Segmentation

Blockwise segmentation is performed by first dividing the image into overlapping blocks, where each block contains m × m pixels, and adjacent blocks overlap by m/2 pixels in both the horizontal and vertical directions. The blocks are denoted by Oi,j for i = 1, .,M, and j = 1, .,N, where M and N are the number of the blocks in the vertical and horizontal directions, respectively. If the height and width of the input image is not divisible by m, the image is padded with zeros. For each block, the color axis having the largest variance over the block is selected and stored in a corresponding gray image block, ˜Oi,j .

The pixels in each block are segmented into foreground ("1") or background ("0") by the clustering method of Cheng and Bouman. The clustering method classifies each pixel in ˜Oi,j by comparing it to a threshold t. This threshold is selected to minimize the total sub-class variance.

B. Global Segmentation

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

The global segmentation step integrates the individual segmentations of each block into a single consistent segmentation of the page. To do this, we allow each block to be modified using a class assignment denoted by, $s_{i,j} \in \{0, 1, 2, 3\}$.

$s_{i,j} = 0 \Rightarrow \tilde{} C_{i,j} = C_{i,j}$ (Original)

$s_{i,j} = 1 \Rightarrow \tilde{} C_{i,j} = \neg C_{i,j}$ (Reversed)

$s_{i,j} = 2 \Rightarrow \tilde{} C_{i,j} = \{0\}_{m \times m}$ (All background)

$s_{i,j} = 3 \Rightarrow \tilde{} C_{i,j} = \{1\}_{m \times m}$ (All foreground)

Notice that for each block, the four possible values of $s_{i,j}$ correspond to four possible changes in the block's segmentation: original, reversed, all background, or all foreground. If the block class is "original", then the original binary segmentation of the block is retained. If the block class is "reversed", then the assignment of each pixel in the block is reversed (i.e. 1 goes to 0, or 0 goes to 1). If the block class is set to "all background" or "all foreground", then the pixels in the block are set to all 0's or all 1's, respectively. Figure 4 illustrates an example of the four possible classes where black indicates a label of "1" (foreground) and white indicates a label of "0" (background). Our objective is then to select the class assignments, $s_{i,j} \in \{0, 1, 2, 3\}$, so that the resulting binary masks, $\tilde{} C_{i,j}$, are consistent.

III. CONNECTED COMPONENT CLASSIFICATION (CCC)

The connected component classification (CCC) algorithm refines the segmentation produced by COS by removing many of the erroneously detected non-text components. The CCC algorithm proceeds in three steps: connected component extraction, component inversion, and component classification. The connected component extraction step identifies all connected components in the COS binary segmentation using a 4-point neighborhood. In this case, connected components less than six pixels were ignored because they are nearly invisible at 300 dpi resolution. The component inversion step corrects text segmentation errors that sometimes occur in COS segmentation when text is locally embedded in a highlighted region (See Fig. 5 (a)). Figure 5 (b) illustrates this type of error where text is initially segmented as background.

Notice the text "100 Years of Engineering Excellence" is initially segmented as background due to the red surrounding region. In order to correct these errors, we first detect foreground components that contain more than eight interior background components (holes). In each case, if the total number of interior background pixels is less than half of the surrounding foreground pixels, the foreground and background assignments are inverted. Figure 5 (c) shows the result of this inversion process. Note that this type of error is a rare occurrence in the COS segmentation.

The final step of component classification is performed by extracting a feature vector for each component, and then computing a MAP estimate of the component label. The feature vector, y_i, is calculated for each connected component, CC_i, in the COS segmentation. Each y_i is a 4 dimensional feature vector which describes aspects of the ith connected component including edge depth and color uniformity.

Finally, the feature vector y_i is used to determine the class label, x_i, which takes a value of 0 for non-text and 1 for text.

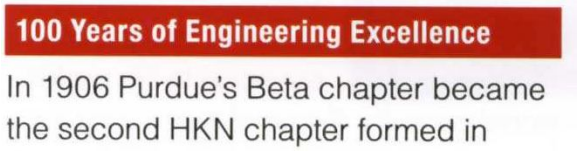

Fig 5(a) Original image

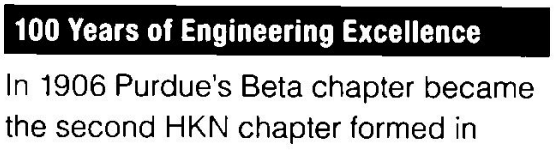

Fig 5(b) Initial segmentation

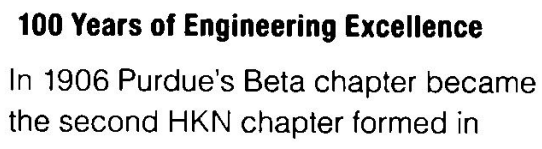

Fig 5(c) Preprocessed segmentation

The conditional distribution of the feature vector y_i given x_i is modeled by a multivariate Gaussian mixture while the underlying true segmentation labels are modeled by a Markov random field (MRF).

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

IV. MATLAB SIMULATION OUTPUT

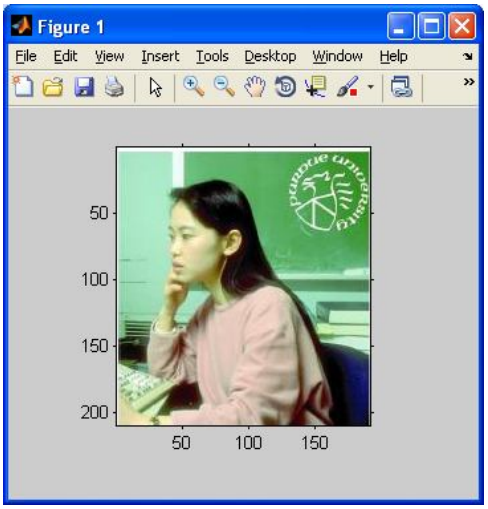

Fig 6 : Input image

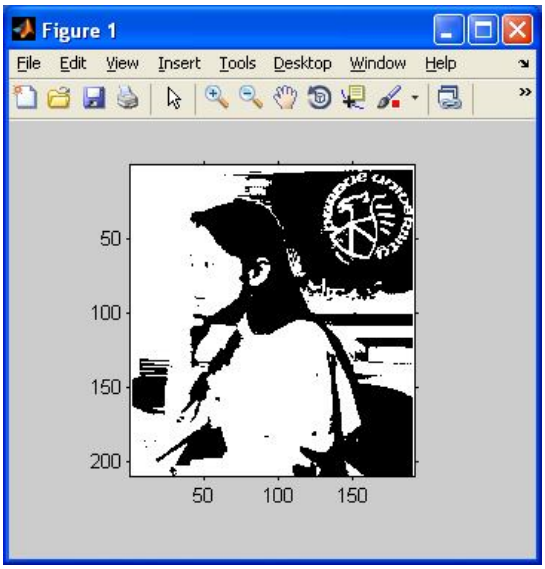

Fig 7 : LURA & COS

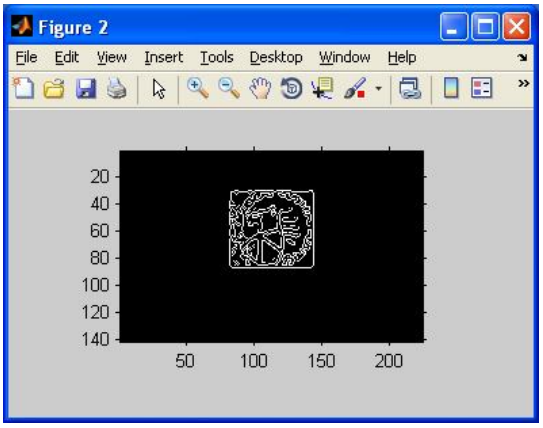

Fig 8: COS/CCC Output

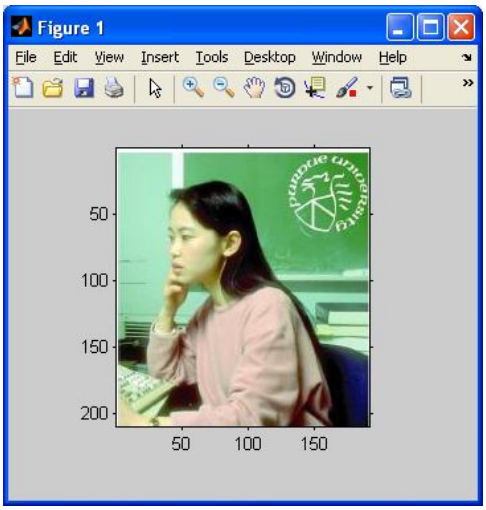

Fig 9: Decompression Image

V.CONCLUSION

We presented a novel segmentation algorithm for the compression of raster documents. While the COS algorithm generates consistent initial segmentations, the CCC algorithm substantially reduces false detections through the use of a component-wise MRF context model. The MRF model uses a pair-wise Gibbs istribution which more heavily weights nearby components with similar features. We showed that the COS/CCC algorithm achieves greater text detection accuracy with a lower false detection rate, as compared to state-of-the-art commercial MRC products. Such text-

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

only segmentations are also potentially useful for document processing applications such as OCR.

REFERENCES

1. International Telecommunication Union, *ITU-T recommendation T.44 Mixed raster content (MRC)*, April 1999.
2. L. Bottou, P. Haffner, P. G. Howard, P. Simard, Y. Bengio, and Y. LeCun, "High quality document image compression with DjVu," *Journal of Electronic Imaging*, vol. 7, no. 3, pp. 410–425, 1998.
3. P. Haffner, L. Bottou, and Y. Lecun, "A general segmentation scheme for DjVu document compression," in *Proc. of ISMM 2002*, Sydney, Australia, April 2002.
4. "Luradocument pdf compressor," available from https://www.luratech.com.
5. Y. Zheng and D. Doermann, "Machine printed text and handwriting identification in noisy document images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 3, pp. 337–353, 2004.
6. S. Kumar, R. Gupta, N. Khanna, S. Chaundhury, and S. D. Joshi, "Text extraction and document image segmentation using matched wavelets and MRF model," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2117–2128, 2007.
7. J. Kuk, N. Cho, and K. Lee, "MAP-MRF approach for binarization of degraded document image," in *Proc. of IEEE Int'l Conf. on Image Proc.*, 2008, pp. 2612–2615.
8. H. Cao and V. Govindaraju, "Processing of low-quality handwritten documents using Markov Random Field," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 7, pp. 1184–1194, 2009.
9. J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th International Conf. on Machine Learning*. Morgan Kaufmann, 2001, pp. 282–289.
10. S. Kumar and M. Hebert, "Discriminative random fields: A discriminative framework for contextual interaction in classification," in *Proc. of Int'l Conference on Computer Vision*, vol. 2, 2003, pp. 1150–1157.
11. M. C.-P. a. X. He, R.S. Zemel, "Multiscale conditional random fields for image labeling," *Proc. of IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 695–702, 2004.
12. M. Li, M. Bai, C. Wang, and B. Xiao, "Conditional random field for text segmentation from images with complex background," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2295–2308, 2010.
13. E. Haneda, J. Yi, and C. A. Bouman, "Segmentation for MRC compression," in *Color Imaging XII: Processing, Hardcopy, and Applications*, vol. 6493, San Jose, CA, 29th January 2007.
14. T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd ed. McGraw-Hill, 2001.
15. J. Besag, "On the statistical analysis of dirty pictures," *J. Roy. Statist. Soc. B*, vol. 48, no. 3, pp. 259–302, 1986.
16. C. A. Bouman, "Digital Image Processing Laboratory: Markov Random Fields and MAP Image Segmentation," January2007, available from http://www.ece.purdue.edu/˜bouman.
17. Ricardo de Queiroz, Robert Buckley and Ming Xu Corporate Research & Technology, Xerox Corp.