DECODING OF TEXT LINES IN GRAYSCALE DOCUMENT IMAGES

Kris Popat

Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304 USA
email: popat@parc.xerox.com

ABSTRACT

The Document Image Decoding (DID) framework for recognizing printed text in images has been shown in previous work to achieve extremely high recognition accuracy when its models are well matched to the data. To date, DID has been restricted to binary images, in part for computational reasons, and in part because binary scanning is widely available and often of sufficient spatial resolution to make the use of grayscale information unnecessary for reliable recognition. Advances in computer speed and memory, along with the emergence of low-cost digital still cameras and similar devices as alternatives to traditional scanners, motivates the extension of the DID formalism to the low-spatial-resolution grayscale and color domains. To do so requires substantially generalizing DID's image-formation and degradation models. This paper lays out an approach and presents preliminary results on real data.

1. INTRODUCTION

Low-cost digital cameras are poised to offer a convenient alternative to flatbed scanners for document image acquisition in certain settings. For example, one might use a pocket digital camera to take snapshots of selected pages while browsing books in a library, rather than having to carry the books over to a scanner or copy machine. The newest digital cameras provide sufficient resolution to allow humans to read the resulting images, but machine-recognition of the text remains challenging because of the geometric distortions, nonuniform illumination, and the severely limited spatial resolution when compared with traditional scanning devices.

Document Image Decoding (DID) [1] is an approach to text recognition based on a communications systems view that has been found to achieve high recognition accuracy when its models are well-matched to the data [2, 3, 4, 5]. To date, work on DID has focused on binary images. To extend it to the low-resolution grayscale or color domains, both its image-formation and degradation models must be extended. This paper proposes modeling the physical processes by *simulation* where possible. The general strategy is to form hypotheses in a high-resolution domain, then evaluate these hypotheses against the observed image after simulating the loss of spatial resolution and the other distortions incurred in the imaging process.

1.1. Text-Line Image Decoding

In the DID framework, document images are regarded as having been produced by transitioning through a Markov source. The source begins in a *start* state and terminates in a *stop* state. Each transition within the source causes the rendering of a character template (a bitmap) on the page at the current cursor position, then advances that position by (in general) a two-dimensional vector displacement in preparation for printing the next character. The set of character templates includes whitespace of various kinds. Formally, each transition in the source is assigned a four-tuple consisting of a character template, the two-dimensional displacement by which to advance the cursor, the prior probability of following that transition, and a string label. Every complete path through the source defines a document image and an associated transcription: the image is the superposition of the bitmaps rendered on each transition, and the transcription is the concatenation of the associated string labels.

After the document image has been formed in this way, it is assumed to be subjected to random corruption, which causes uncertainty in the recognition process. Recognition involves finding a complete path through the hypothesized Markov source that best explains the observed image. In particular, a complete path is sought that is *a posteriori* most probable considering the entire image as evidence, where the probability is computed on the basis of source and degradation models. Finding the most probable path is not the same as finding the most probable message (at issue is the well-known *Viterbi approximation* [6]), but finding the most probable path is simpler to do and experience has shown that it nevertheless results in accurate recognition.

When DID is applied to a single line of text instead of to

an arbitrary page, a Markov source with a minimal structure can be used. Specifically, it can consist of a start state, a single interior state, and a stop state. The interior state has one self-transition for each character template in the font. Generation of a text line begins in the start state, with the cursor at the leftmost horizontal position in the text-line image. The first transition is into the interior state, and subsequent transitions loop back into that state, each time imaging a character and advancing the cursor horizontally. After the text line has thus been produced, a final transition is made into the stop state at the rightmost position in the text-line image, whereupon the process terminates.

During the recognition process, a *score* is associated with each transition along a candidate path. This score accounts for both the prior probability of following the transition, and the likelihood of the transition given the segment of the observed text line image defined by the horizontal positions before and after the transition. The likelihood term is the one of interest here, as it is the link between the observed image and the decoding operation.

2. LOW-RESOLUTION GRAYSCALE DID

In principle, extending DID to the grayscale or color domains can be accomplished simply by defining an appropriate likelihood function for matching hypothesized characters against the image at all feasible positions. A source of complexity is that multiple imaged instances of a character exhibit a systematic variation of the edge pixel values accruing from spatial sampling, and this variation is not well described by an independent additive noise degradation model traditionally assumed in DID [1, 2]. A direct application of the traditional apparatus would necessitate the association of multiple templates with each character, to reflect and accommodate the systematic variability due to the relatively coarse sampling.

Instead we describe an alternative approach in which the hypothesis search is carried out using single templates for each character but in a high-spatial-resolution domain. The blur, sampling, and degradation processes are numerically simulated to provide a likelihood function for each template at a lower-resolution, which can then be evaluated against the observed grayscale image after appropriate displacement. There are restrictions on the types of degradation that can be modeled by this approach, but these are considerably weaker than in previous DID formulations.

2.1. Image Formation Model

We observe an $n_1 \times n_2$ grayscale image z, which we regard as a degraded, low-resolution version of a hypothetical $N_1 \times N_2$ image Z. These two images are related by the composition of transformations shown in Figure 1. Pixel

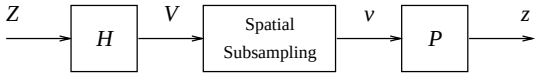

Fig. 1. A model of grayscale document image formation.

values in both Z and z are assumed to take on scalar values in a finite range. For concreteness we will assume that this range is $[0, \ldots, 255]$ for both images. We assume that the spatial subsampling process consists of retaining only those pixels whose first coordinate is a multiple of an integer M_1 and whose second coordinate is a multiple of an integer M_2. Accordingly, the dimensions of Z and z are related by $n_1 = \lfloor N_1/M_1 \rfloor$ and $n_2 = \lfloor N_2/M_2 \rfloor$, where $\lfloor x \rfloor$ denotes the greatest integer not greater than x. It is assumed that H is a spatially local, possibly nonlinear transformation, and that it preserves any image consisting entirely of zero-valued pixels. The transformation P is assumed to corrupt each pixel in its input according to a probability law that is independent of both the corruption made to input pixels in other positions and of their uncorrupted values, but which may depend on spatial position and on the uncorrupted pixel value in the same position. In the simplest case H is a linear blur operator, and P consists of adding independent, identically-distributed noise followed by quantizing back into the range $[0, \ldots, 255]$.

We view the uncorrupted image Z as having been formed in the usual DID way. Specifically, a path through a Markov source defines a sequence of templates, each positioned at a point in Z subject to the requirement that no two templates overlap in their non-zero-valued pixels. Note that these high-resolution templates in Z may be either binary or gray-valued; our analysis will hold in either case. However, in view of the subsequent application of the transformation H in the model and the smoothing effect this can have on edge pixels, is should be noted that realism is not substantially sacrificed by assuming the hypothetical high-resolution templates to be binary.

2.2. Match-Score Computation

To apply the DID recognition framework we must have a way of computing the individual contribution made by each imaged template to the overall likelihood, without knowledge of what other templates might have been imaged elsewhere along the same path (this condition is to make the search tractable by known algorithms [7]). Established DID methodology tell us how to accomplish this when a partially bit-flip-corrupted version of Z is taken as the input. We seek here instead to allow z to be the input, taking into account in the computation the more complicated distortions accruing from the composition of the blur H, subsampling, and

signal-dependent spatially varying point-transformation P.

Note that the composition of H and subsampling is a periodically spatially varying, possibly nonlinear, but purely deterministic process which when applied to Z results in a deterministic image $v(Z)$. We can therefore identify the overall likelihood with that of transforming $v(Z)$ into z. Moreover, P is a purely random transformation in which, for any given input, the output pixels are conditionally independent of one another. We can characterize P by a collection of input-output distributions $p_{i,j}(\cdot\,|v_{i,j}(Z))$, one for each pixel position (i,j) in z. The log likelihood is then

$$l(z|Z) = \sum_{i,j} \log p_{i,j}(z_{i,j}|v_{i,j}(Z)) \tag{1}$$

where the summation is taken over all pixel positions in z, and where $z_{i,j}$ is the value of z at (i,j).

The next task is to break (1) apart in terms of the templates imaged in Z. Let $C_{x,y}$ denote the event "template C has been placed at position (x,y) in Z." Let $S_Z(C_{x,y})$ denote the *support* of $C_{x,y}$ in Z; that is, the set of coordinates of those pixels in Z that are nonzero as a result of the event $C_{x,y}$. Let $S_V(C_{x,y})$ and $S_v(C_{x,y})$ denote the corresponding support in V and v, respectively. We can now be precise about the locality requirement on the blur process: H must be such that $S_V(C_{x,y})$ and $S_V(C_{x'y'})$ are disjoint whenever $S_Z(C_{x,y})$ and $S_Z(C_{x'y'})$ are. (Recall that a separate assumption requires the latter condition to hold whenever $C_{x,y}$ and $C_{x'y'}$ occur on the same path.) Since subsampling results in at most a subset of pixels being retained, $S_V(C_{x,y}) \cap S_V(C_{x'y'}) = \emptyset$ implies that $S_v(C_{x,y}) \cap S_v(C_{x'y'}) = \emptyset$. This allows us to conclude that those nonzero pixels in v caused by $C_{x,y}$ would have been zero had $C_{x,y}$ not occurred on the path, all else being the same. We can therefore rewrite (1) as

$$\begin{aligned} l(z|Z) &= \sum_{i,j} \log p_{i,j}(z_{i,j}|v_{i,j}(Z_0)) \\ &+ \sum_{C_{x,y}} \sum_{(i,j)\in S_v(C_{x,y})} \log \frac{p_{i,j}(z_{i,j}|v_{i,j}(Z|C_{x,y}))}{p_{i,j}(z_{i,j}|v_{i,j}(Z_0))} \end{aligned} \tag{2}$$

where Z_0 is an all-zero image of the same dimensions as Z, where the summation over $C_{x,y}$ is understood to be over those templates imaged by the Markov source when Z was generated, and where $v_{i,j}(Z|C_{x,y}))$ denotes the pixel in position (i,j) of v given that $C_{x,y}$ occurred in generating Z. Since the first term in (2) is independent of the hypothesis image Z, it can be omitted when using the expression to judge the degree of match between Z and z. We therefore define an overall match score as

$$\text{match}(z|Z) = \sum_{C_{x,y}} \text{match}(z|C_{x,y}) \tag{3}$$

where the individual contribution of each template is defined as

$$\text{match}(z|C_{x,y}) = \sum_{(i,j)\in S_v(C_{x,y})} \log \frac{p_{i,j}(z_{i,j}|v_{i,j}(Z|C_{x,y}))}{p_{i,j}(z_{i,j}|v_{i,j}(Z_0))} \tag{4}$$

Expression (4) provides a match score that can be used to label edges in a DID trellis. To compute it, we note that $v_{i,j}(Z|C_{x,y})$ is periodic in the sense that

$$v_{i,j}(Z|C_{x,y}) = v_{i-\lfloor x/M_1 \rfloor, j-\lfloor y/M_2 \rfloor}(Z|C_{x \bmod M_1, y \bmod M_2}) \tag{5}$$

and its support is likewise periodic. Thus, every low resolution imaged template in v can be represented as a translation of one of at most $M_1 \times M_2$ distinct patterns, each corresponding to a distinct subsampling phase. These patterns and their corresponding supports can be pre-computed and stored in a table, then recalled for use in (4) when it is required to score the match of a hypothesized character at a particular position (x,y) in Z. Recent advances in the use of heuristic match scores in DID [8] reduce the importance of (4) being very fast to compute, provided that a suitable, computationally inexpensive heuristic upper bound on (4) can be found.

The remaining issue in applying (4) is the estimation and evaluation of p. The dependence of p on (i,j) in (4) provides a means of incorporating into the model nonstationary phenomena such as spatially varying illumination. Once the structure for p has been specified, its remaining parameters can be learned from example data.

3. EXPERIMENTAL RESULTS

A 1024×768, eight-bit-per-pixel test image was obtained using a handheld Sony DSC-F505 digital camera about 60 cm above a deliberately wrinkled, 15-line test document set in a known font. The picture was taken under low intensity oblique lighting. Although the lighting conditions and resolution were chosen to be challenging, care was taken to avoid any significant geometric distortions. After scanning, the image was cropped, and a global rotation correction was applied via bilinear interpolation. A fragment is shown in Figure 2; two complete lines in Figure 3. The model transformation H was manually chosen to consist of a morphological erosion using a 3×3 square structuring element, followed by a separable lowpass filter. A suitable subsampling factor was empirically determined to be 2 in both dimensions. The random corruption was modeled as a spatially varying gain operator (accounting for the nonuniform illumination) followed by additive zero-mean Gaussian noise, and finally re-quantization into the range $[0,\ldots,255]$. For simplicity, the variance of each added noise value was assumed to depend only on the gain-normalized uncorrupted

pixel value in the same location; the relationship was estimated using local averages to approximate the uncorrupted pixels. The local illumination estimate for the deterministic, spatially varying part of P was obtained at each pixel location by computing the top quartile value along a 60-pixel-long line segment centered on the current pixel and oriented to yield minimum intensity variance along the segment. Text baselines were identified in the image by detecting peaks among pairwise differences of pixel sums taken along adjacent rows. A single-space null-string transition was included among the candidate transitions to allow fine-spacing alignment. Each candidate match was repeated at vertical positions one-pixel above and one-pixel below the assigned baseline, to accommodate slight deviations from baseline linearity. A unigram language model was used to provide prior weights for the trellis edges. The observed number of recognition errors for the 963-character test document were: 55 substitutions, 15 deletions (mostly spaces and punctuation), and 0 insertions, as determined by a dynamic programming text alignment procedure. Considering the high noise levels in the image and the severely nonuniform lighting conditions, these results are felt to be encouraging. Additional details about the experiment are available at http://www.parc.xerox.com/popat/graydidicassp.html.

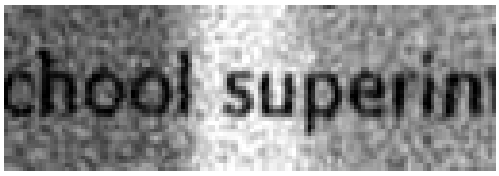

Fig. 2. A fragment of the test image. Note limited spatial resolution and high sensor noise due to low-light conditions.

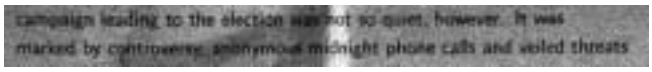

Fig. 3. Two text lines from the 15-line test image.

4. CONCLUSION

We have extended the Document Image Decoding framework to function in the low-resolution grayscale domain. The work is important because of the emergence of low-cost digital cameras as alternative and attractive input devices for document-image capture. The extension relies in large part on numerical simulation of the imaging process; it is based on performing the search in an idealized hypothesis image space, while evaluating the hypotheses (i.e., computing the likelihood) in the low-resolution, degraded, observation image space. The expanded framework accommodates realistic types of distortion. Preliminary results have been encouraging and serve as a proof-of-principle. More work is required to automate the inference of the blur, subsampling, and font parameters; to accommodate and correct for projective and other geometric distortions; to apply more sophisticated language models [7]; and to assess performance by conducting larger scale experiments.

5. REFERENCES

[1] Gary E. Kopec and Philip A. Chou, "Document image decoding using Markov source models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, pp. 602–617, June 1994.

[2] Gary E. Kopec, "Multilevel character templates for document image decoding," in *Proceedings of the IS&T/SPIE 1997 Intl. Symposium on Electronic Imaging: Science & Technology*, San Jose, February 1997.

[3] Gary E. Kopec and Phil A. Chou, "Markov source model for printed music decoding," in *Proceedings of the SPIE Document Recognition II Conference*. SPIE–the International Society for Optical Engineering, February 1995, vol. 2422, pp. 115–125, SPIE.

[4] Gary E. Kopec, "Document image decoding in the uc berkeley digital library," in *Proceedings of the SPIE Document Recognition III Conference*. SPIE–the International Society for Optical Engineering, January 1996, vol. 2660, pp. 2–13, SPIE.

[5] Gary E. Kopec, Philip A. Chou, and David A. Maltz, "Markov source model for printed music decoding," *Journal of Electronic Imaging*, vol. 5, no. 1, pp. 7–14, January 1996.

[6] Frederick Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, Massachusetts, 1997.

[7] Kris Popat, Dan Greene, Justin Romberg, and Dan S. Bloomberg, "Adding linguistic constraints to document image decoding: Comparing the iterated complete path and stack algorithms," in *Proceedings of IS&T/SPIE Electronic Imaging 2001: Document Recognition and Retrieval VIII*, January 2001.

[8] Thomas P. Minka, Dan S. Bloomberg, and Kris Popat, "Document image decoding using iterated complete path heuristic," in *Proceedings of IS&T/SPIE Electronic Imaging 2001: Document Recognition and Retrieval VIII*, January 2001.