Multilevel Character Templates for Document Image Decoding*

Gary E. Kopec
Xerox Palo Alto Research Center
kopec@parc.xerox.com

ABSTRACT

Early work in document image decoding (DID) was based on a bilevel imaging model in which an observed image is formed by passing an ideal bilevel image through a memoryless asymmetric bit-flip channel. While this simple model has proven useful in practice, there are many situations in which the bit-flip channel is an inadequate degradation model. This paper presents a multilevel generalization of the bilevel model in which the pixels of the ideal image are assigned values from a finite set of L discrete "colors" or levels. Level 0 is a background color and the remaining levels are foreground colors. The observed image is bilevel and is modelled as the output of a memoryless L-input symbol, 2-output symbol channel. The multilevel model is motivated in part by the intuition that pixels in a character image are more or less reliably black, depending on their distance from an edge. In addition, the multilevel model supports both "write-black" and "write-white" levels, and thus can be used to implement a probabilistic analog of morphological "hit-miss" filtering. In experiments with the Univ. of Washington UW-II English journal database, the character error rate with multilevel templates ($L = 4$) was about a factor of four less than the error rate with bilevel templates.

1. INTRODUCTION

Document image decoding (DID) is a methodology for formulating, analyzing and solving document recognition problems that is based on the communication theory view of document creation, transmission and interpretation shown in fig. 1 [5]. A stochastic *message source* selects a finite string M from a set of candidate strings according to a prior probability distribution. The *imager* converts the message into an ideal image Q. Collectively, the message source and imager form the *image source*. The *channel* maps the ideal image into an observed image Z by introducing distortions due to printing and scanning, such as skew, blur and additive noise. Finally, the *decoder* receives image Z and produces an estimate $\hat{M}$ of the original message, according to a maximum a posteriori (MAP) decision criterion.

Initial investigations of DID focussed on *bilevel* imaging scenarios in which Q and Z are binary images and pixel values

*Presented at *IS&T/SPIE 1997 Intl. Symposium on Electronic Imaging: Science & Technology*, San Jose, CA, Feb. 8–14, 1997.

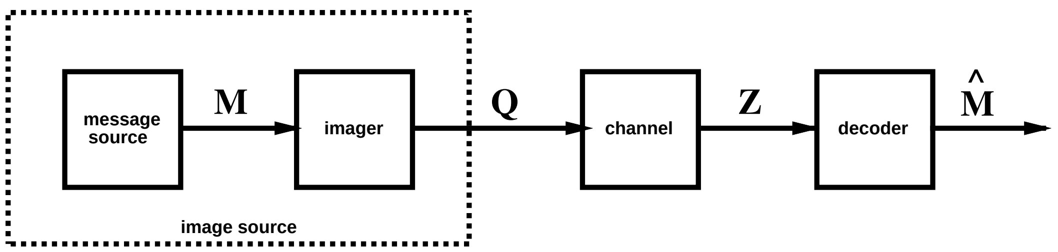


Figure 1: Communication theory view of document recognition.

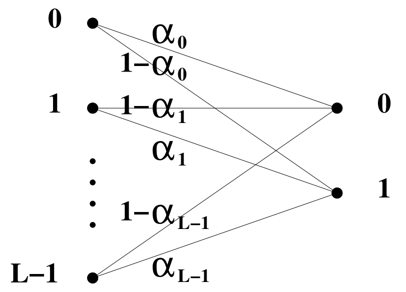


Figure 2: L-input symbol, 2-output symbol discrete memoryless channel model.

of 0 and 1 correspond to background (white) and foreground (black), respectively [5]. The standard bilevel channel is the memoryless asymmetric bit-flip channel characterized by two parameters

$$\begin{aligned} \alpha_0 &= \Pr\{z(\vec{x}) = 0 \mid q(\vec{x}) = 0\} \\ \alpha_1 &= \Pr\{z(\vec{x}) = 1 \mid q(\vec{x}) = 1\}. \end{aligned}$$

The bilevel model is analytically simple and intuitively straightforward. The pixel values of Q can be directly interpetted as black and white image colors and the channel represents a simple "salt-and-pepper" noise process. Moreover, the bilevel model has been found to work well in practice, even in many situations in which the actual image degradation is not salt-and-pepper noise. However, the simple bit-flip channel is clearly inadequate as a general image degradation model and thus the development of more sophisticated imaging models for DID is an important research problem.

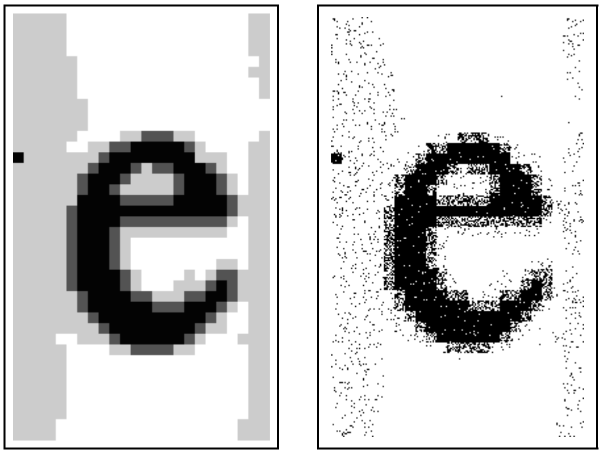
A natural generalization of the bit-flip image model is a *multilevel* imaging model, in which the pixel values of Q are drawn from a finite set of *colors* (or *levels*), i.e. $q(\vec{x}) \in \{0, \ldots, L-1\}$. Level 0 denotes the background color and levels $1, \ldots, L-1$ are foreground colors. The observed image Z is bilevel, as before. The multilevel channel is the L-input symbol, 2-output symbol discrete memoryless channel shown in fig. 2. The multilevel channel parameters are defined so that the case $L = 2$ corresponds exactly to the bilevel model.

The multilevel model is motivated, in part, by the common observation that pixels in a character image are more or less "reliable", depending on their distance from an edge. In a multilevel model, different levels might correspond to "black pixel near edge", "black pixel away from edge", "white pixel near edge", etc. This is illustrated in fig. 3, which shows a four-level ideal image (or template) Q_e for the letter 'e' and a bilevel observed image Z created from Q_e. The levels in fig. 3(a) are indicated using shades of gray whose darkness increases with the probability that $z(\vec{x}) = 1$, for a channel with parameters $\alpha_0 = .95$, $\alpha_1 = .98$, $\alpha_2 = .0001$ and $\alpha_3 = .5$. The observed image in fig. 3(b) was generated from this ideal image and channel.

The remainder of this paper presents multilevel generalizations of basic aspects of DID theory and discusses a set of experiments to assess the utility of multilevel models in decoding. Sections 2, 3 and 4 discuss multilevel image sources, channels and decoders, respectively. Section 5 outlines multilevel extensions to the DID approach to character template estimation presented in [6]. Finally, section 6 describes an evaluation of document-specific multilevel templates using the English journal subset of the Univ. of Washington UW-II database [9]. Use of four-level templates reduced the average character error rate by about a factor of four compared with bilevel templates.

2. MULTILEVEL IMAGE SOURCE

A multilevel image source defines the universe of ideal images Q for a given recognition problem and specifies a probability distribution on them. As in bilevel DID, a useful class of multilevel image sources are the *Markov sources*, which use finite-state grammars to describe how composite ideal images are constructed from a set of primitive multilevel "templates". Formally, multilevel and bilevel Markov image sources are nearly identical.



(a) (b)

Figure 3: Multilevel imaging example. (a) Four-level ideal image (template) of 'e'. Light gray: level 0, black: level 1, white: level 2, medium gray: level 3. (b) Bilevel observed image of 'e'. Channel parameters: $\alpha_0 = .95$, $\alpha_1 = .98$, $\alpha_2 = .0001$ and $\alpha_3 = .5$.

A multilevel Markov image source is a directed graph consisting of a finite set of states (nodes, vertices) $\mathcal{N}$ and a set of directed transitions (branches, edges) $\mathcal{B}$ [5]. Two distinguished members of $\mathcal{N}$ are the initial state n_I and the final state n_F. Each transition t connects a pair of states, L_t and R_t, that are called, respectively, the *predecessor* (left) state and the *successor* (right) state of t. Each transition is labeled with a 4-tuple of attributes, $(Q_t, m_t, a_t, \vec{\Delta}_t)$, where Q_t is a multilevel *template*, m_t is a *message* string, a_t is the *transition probability* and $\vec{\Delta}_t$ is the vector *displacement* of t.

The pixel value (color) of template Q_t at position $\vec{x}$ relative to the local origin of the template is denoted by $q_t(\vec{x})$. The "color" of a multilevel template pixel does not, in general, have an immediate physical interpretation as an intensity or color in the ordinary sense (e.g. RGB coordinates). Rather, the color of $q(\vec{x})$ is simply an identifier that specifies, via the channel, a probability distribution on the values of the corresponding observed image pixel $z(\vec{x})$.

In principle, each template extends over the entire image plane. However, the *support* of Q_t (the set of locations where $q_t(\vec{x})$ differs from the background color, i.e. 0) is typically localized to a small region in the vicinity of the origin. In particular, it is typically assumed that the support of each template lies within a rectangle $\mathcal{C}$, called the *canvas* of the template.

A *complete path* Π through a Markov source is a sequence of transitions $t_0, \ldots, t_{P-1}$ where $L_{t_0} = n_I$, $R_{t_{P-1}} = n_F$ and $R_{t_{i-1}} = L_{t_i}$ for $i = 1, \ldots, P-1$. A complete path defines a sequence of positions in the image plane, $\vec{x}_0, \ldots, \vec{x}_P$, recursively defined by $\vec{x}_{i+1} = \vec{x}_i + \vec{\Delta}_{t_i}$, where $\vec{x}_0$ is an initial position, normally $\vec{0}$. These image positions define a composite ideal image Q_Π by

$$Q_\Pi = \biguplus_{i=0}^{P-1} Q_{t_i}[\vec{x}_i] \tag{1}$$

where $Q_{t_i}[\vec{x}_i]$ denotes Q_{t_i} shifted so that the origin of its local coordinate system is located at $\vec{x}_i$ and $\biguplus$ represents a pixel color composition rule that defines how the colors of overlaping template pixels combine to give the color of the composite image pixel. The pairs $\{(\vec{x}_i, t_i), i = 0, \ldots, P-1\}$ are called the *labeled glyph positions* defined by Π and we will let $\vec{x}_1^{(t)}, \ldots, \vec{x}_{N_t}^{(t)}$ denote the coordinates of the glyph positions labeled with a particular transition t, where N_t is the number of times transition t occurs in Π.

Markov image sources are normally designed to satisfy a *template disjointness constraint* that asserts that the shifted tem-

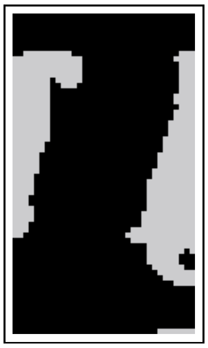


Figure 4: Template for italic space character. Light gray: level 0 ($\alpha_0 = .95$), black: level 1 ($\alpha_1 = .0001$).

plates $Q_{t_i}[\vec{x}_i]$ on the right hand side of (1) have disjoint supports. Because of template disjointness, $\uplus$ may be taken as simple pixelwise addition. In the case of bilevel images this reduces to the bitwise **or** operation.

3. MULTILEVEL CHANNEL

The multilevel channel defines a set of probability distributions for the observed image pixel values, one distribution for each template pixel color. For bilevel observed images Z, these distributions are characterized by the level parameters α_i.

A foreground level l (i.e. $l > 0$) has one of two physical interpretations, depending on the relationship between α_l and α_0. If $\alpha_l > 1 - \alpha_0$ then a pixel on level l is more likely than a background pixel to be observed as black. This corresponds to the common situation of black-on-white, or "write-black", printing. Conversely, if $\alpha_l < 1 - \alpha_0$ then a level l pixel is less likely than a background pixel to be observed black. This corresponds, loosely, to "reverse video", or "write-white" printing.

A given template (or set of templates) may include both write-black and write-white levels. In fig. 3, for example, level 2 (white) is write-white while levels 3 (medium gray) and 1 (black) are write-black. Templates with both write-black and write-white levels are similar to the hit-miss transforms used for morphological shape recognition [2].

An important use for write-white levels is in "space" character templates. Modeling space characters as templates allows their parameters to be estimated in the same way as those of normal "printing" characters. Fig. 4 shows a template for a space character in an italic font. The corresponding channel parameters are $\alpha_0 = .95$ and $\alpha_1 = .0001$. Level 1 is write-white since $\alpha_1 < 1 - \alpha_0$.

The formal role of a channel model in DID is to provide an expression for the log likelihood ratio

$$\mathcal{L}(Z \mid Q) \equiv \log \tilde{\Pr}\{Z \mid Q\} \tag{2}$$

$$\equiv \log \frac{\Pr\{Z \mid Q\}}{\Pr\{Z \mid Q_0\}} \tag{3}$$

where Q_0 is the all-background image [5]. For the bilevel source and channel this reduces to

$$\mathcal{L}(Z \mid Q) = \gamma \, ||Q \wedge Z|| + \beta \, ||Q|| \tag{4}$$

where $||X||$ denotes the number of 1's in X, $\wedge$ is the bitwise **and** operator and

$$\gamma = \log \frac{\alpha_0 \alpha_1}{(1-\alpha_0)(1-\alpha_1)} \tag{5}$$

$$\beta = \log \frac{1-\alpha_1}{\alpha_0} \tag{6}$$

as derived in [5].

The derivation of (4) can be generalized to multilevel models by treating each foreground level as a separate bilevel template. If Q is a multilevel template, let $\tilde{Q}^{(l)}$ be the bilevel bitmap representing the l^{th} level set of Q, defined by

$$\tilde{q}^{(l)}(\vec{x}) \equiv \begin{cases} 1 & \text{if } q(\vec{x}) = l \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Then, it is not difficult to show that

$$\mathcal{L}(Z \mid Q) = \sum_{l=1}^{L-1} \left[\gamma_l \,\|\tilde{Q}^{(l)} \wedge Z\| \;+\; \beta_l \,\|\tilde{Q}^{(l)}\|\right] \tag{8}$$

$$= \sum_{l=1}^{L-1} \mathcal{L}(Z \mid \tilde{Q}^{(l)}) \tag{9}$$

where γ_l and β_l are given by (5) and (6) for each level. For a write-white level, $\gamma_l < 0$ and $\beta_l > 0$. As a result, $\mathcal{L}(Z \mid \tilde{Q}^{(l)})$ will be maximally positive when $\tilde{Q}^{(l)}$ is positioned over pure whitespace and will decrease as more black pixels fall under $\tilde{Q}^{(l)}$.

The significance of (9) is that many properties of bilevel templates extend naturally to the multilevel case. For example, if Q_1 and Q_2 are multilevel templates with disjoint support then

$$\mathcal{L}(Z \mid Q_1 \uplus Q_2) = \mathcal{L}(Z \mid Q_1) + \mathcal{L}(Z \mid Q_2) \tag{10}$$

where $\uplus$ denotes elementwise addition. This follows from (9) and the analogous result for bilevel templates. As a result, from the template disjointness assumption and (1) we have

$$\mathcal{L}(Z \mid Q_\Pi) = \sum_{i=0}^{P-1} \mathcal{L}(Z \mid Q_{t_i}[\vec{x}_i]) \tag{11}$$

which again generalizes a bilevel result.

As noted previously, the bilevel channel has a simple physical interpretation as a salt-and-pepper noise process applied to an ideal bilevel image. While multilevel image sources and channels are straightforward mathematical generalizations of bilevel ones, the correspondence between a multilevel model and any physical imaging process may not be clear. In fact, the multilevel model can be used to represent the statistical properties of many types of local degradation processes. Fig. 5 depicts one such class of degradations which includes the *blur*, *threshold* and *sensitivity* defects defined in [1]. The ideal bilevel image $q(x, y)$ is blurred using a non-negative space-invariant point-spread function $h(x, y)$. The blurred gray-valued image $u(x, y)$ is then corrupted by additive white noise $n(x, y)$ and thresholded to give the observed bilevel image $z(x, y)$. Formally,

$$z(x, y) = \begin{cases} 1 & \text{if } v(x, y) \geq \theta \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

$$v(x, y) = u(x, y) \;+\; n(x, y) \tag{13}$$

$$u(x, y) = \sum h(r, s) q(x - r, y - s) \tag{14}$$

where the sum in (14) is taken over the support of $h(r, s)$. If the noise process $n(x, y)$ has cummulative distribution function $F(t)$, i.e. $F(t) = \Pr\{n(x, y) \leq t\}$, then

$$\Pr\{z(x, y) = 1\} = \Pr\{v(x, y) \geq \theta\} \tag{15}$$

$$= \Pr\{n(x, y) \;\geq\; \theta - u(x, y)\} \tag{16}$$

$$= 1 - F(\theta - u(x, y)). \tag{17}$$

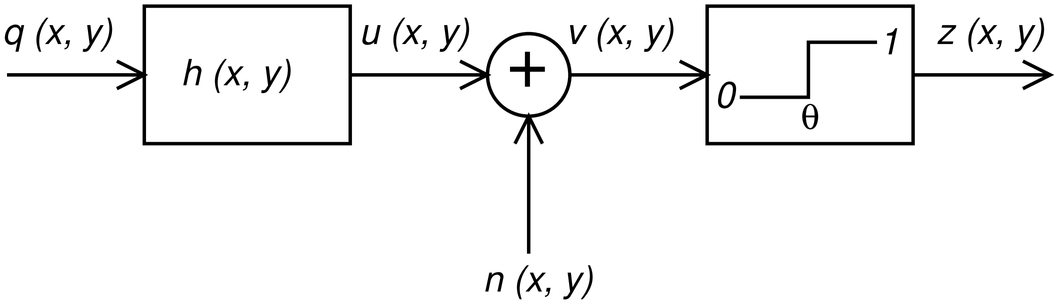


Figure 5: Bilevel image degradation process consisting of blur, additive noise and thresholding.

Constructing a multilevel representation of the process in fig. 5 involves defining the channel parameters $\alpha_0 \ldots \alpha_{L-1}$ and assigning multilevel pixel values $q_m(x,y) \in \{0,\ldots,L-1\}$ so that

$$\Pr\{z(x,y)=1\} \quad = \quad \alpha_{q_m(x,y)}. \tag{18}$$

From (17) it follows that the desired channel parameters and level assignments are given by

$$\alpha_{q_m(x,y)} \quad = \quad 1 - F(\theta - u(x,y)) \tag{19}$$

which leads to a simple procedure for constructing them. Let $u_0 \ldots u_{L-1}$ be the distinct values of the blurred image $u(x,y)$ in increasing order, where it is assumed that $u_0 = 0$. Then it is not hard to see that channel parameters defined by

$$\alpha_i \quad = \quad 1 - F(\theta - u_i), \quad i = 0 \ldots L-1 \tag{20}$$

and color assignments given by

$$q_m(x,y) \quad = \quad l \quad \text{if } u(x,y) = u_l. \tag{21}$$

satisfy (19).

The practicality of the above construction depends on the number of distinct gray values in the blurred image $u(x,y)$. For an optimized optical system in which there is zero cross-talk between non-8-connected pixels, the support of $h(x,y)$ is a 3×3 square centered at the origin [1]. Thus, there are at most 2^9 distinct values of $u(x,y)$. However, if $h(x,y)$ is circularly symmetric, $u(x_o,y_o)$ depends only on the number of 1-valued pixels of $q(x,y)$ at each distance from (x_o,y_o). For a 3×3 square support there are three different distances, containing 1, 4 and 4 pixels, respectively. Thus, the number of levels is at most 50. This upper bound can be further reduced by noting that not all 3×3 patterns are equally likely to occur in ideal text images. Thus, by viewing level selection as a quantization problem for $F(t)$, the number of levels can be traded off against the fidelity of the approximation. As demonstrated below, satisfactory decoding performance can be achieved using only 3 or 4 levels.

4. DECODER

Image decoding using Markov source models with bilevel templates is described in [5]. The kernel of the decoding operation is the recursive computation of the function

$$\mathcal{F}(n;\vec{x}) \;=\; \max_{t|R_t=n} \left\{ \mathcal{F}(L_t;\vec{x}-\vec{\Delta}_t) \;+\; \log a_t \;+\; \mathcal{L}(Z \mid Q_t[\vec{x}-\vec{\Delta}_t]) \right\} \tag{22}$$

for each node n and image position $\vec{x}$. The only change introduced by multilevel templates is in the computation of the template match scores $\mathcal{L}(Z \mid Q_t[\vec{x}])$. However, by (9), if a multilevel template is represented by its level set bitmaps then algorithms for the bilevel template matching operation (4) can be used to compute the multilevel likelihood. Multilevel template matching based on (9) typically requires about $L-1$ times as much computation as bilevel template matching.

5. TEMPLATE ESTIMATION

One of the important practical advantages of the overall DID methodology is a maximum likelihood approach to supervised template estimation from unsegmented page images and unaligned training transcriptions [6]. The central step in template estimation is called *aligned template estimation* (ATE) and involves finding a set of values for $q_t(\vec{x}), \forall t \in \mathcal{B}, \forall \vec{x} \in \mathcal{C}$, that maximizes $\mathcal{L}(Z \mid Q_\Pi)$ for a given training image Z and alignment path Π (equivalently, a set of labeled glyph positions) while satisfying the template disjointness constraint. The ATE algorithm developed in [6] for bilevel templates can be easily modified for the multilevel case.

By expanding the right hand side of (8) in terms of individual template pixels, $\mathcal{L}(Z \mid Q_{t_i}[\vec{x}_i])$ in (11) can be written

$$\mathcal{L}(Z \mid Q_{t_i}[\vec{x}_i]) = \sum_{\vec{x} \in \mathcal{C}} \sum_{l=1}^{L-1} \left[\gamma_l \tilde{q}_{t_i}^{(l)}(\vec{x}) z(\vec{x}+\vec{x}_i) + \beta_l \tilde{q}_{t_i}^{(l)}(\vec{x}) \right]. \tag{23}$$

Substituting this into (11) gives

$$\mathcal{L}(Z \mid Q_\Pi) = \sum_{i=0}^{P-1} \sum_{\vec{x} \in \mathcal{C}} \sum_{l=1}^{L-1} \left[\gamma_l \tilde{q}_{t_i}^{(l)}(\vec{x}) z(\vec{x}+\vec{x}_i) + \beta_l \tilde{q}_{t_i}^{(l)}(\vec{x}) \right]. \tag{24}$$

If the outer summation over path transitions is replaced by a summation over source model transitions, (24) becomes

$$\begin{aligned} \mathcal{L}(Z \mid Q_\Pi) &= \sum_{t \in \mathcal{B}} \sum_{i=1}^{N_t} \sum_{\vec{x} \in \mathcal{C}} \sum_{l=1}^{L-1} \left[\gamma_l \tilde{q}_t^{(l)}(\vec{x}) z(\vec{x}+\vec{x}_i^{(t)}) + \beta_l \tilde{q}_t^{(l)}(\vec{x}) \right] && (25) \\ &= \sum_{t \in \mathcal{B}} \sum_{\vec{x} \in \mathcal{C}} \sum_{l=1}^{L-1} \tilde{q}_t^{(l)}(\vec{x}) \left[\gamma_l \sum_{i=1}^{N_t} z(\vec{x}+\vec{x}_i^{(t)}) + \beta_l N_t \right]. && (26) \end{aligned}$$

In the $L=2$ case of bilevel templates, (26) reduces to

$$\mathcal{L}(Z \mid Q_\Pi) = \sum_{t \in \mathcal{B}} \sum_{\vec{x} \in \mathcal{C}} q_t(\vec{x}) \left[\gamma \sum_{i=1}^{N_t} z(\vec{x}+\vec{x}_i^{(t)}) + \beta N_t \right] \tag{27}$$

which is (5) in [6].

Comparison of (26) and (27) suggests that multilevel estimation is equivalent to bilevel estimation of the level set bitmaps, with two minor differences. First, the level set bitmaps must satisfy the constraint that only one value of $\tilde{q}_t^{(l)}(\vec{x})$ may be non-zero for a given t and $\vec{x}$. However, this will happen automatically as a consequence of the template disjointness constraint. Second, in (27) the same values of γ and β are used for all templates, whereas in (26) they depend on the level. However, this is simply a bookkeeping issue that introduces no real difficulty. Thus, bilevel template estimation algorithms such as those presented in [6] can be easily modified for the general multilevel case.

6. EXPERIMENTAL RESULTS

The utility of multilevel templates for decoding was investigated in a *document-specific decoding* [6, 7, 8] scenario using the English journal subset of the Univ. of Washington UW-II database [9].

The UW-II English journal database contains 622 bilevel scanned images of 311 pages from 30 technical journal articles. Two images of each page are provided, scanned from first and second generation photocopies, respectively. The scanned page images contain approximately 10,400 zones. Each zone is accompanied by a collection of manually prepared records that specify attributes such as zone bounding box coordinates, ground truth text, dominant font size, reading direction, etc. A subset of the zones was used, consisting of those with attribute values

`ZONE_CONTENT`	=	text
`ZONE_TYPE`	=	text-body
`CHARACTER_ORIENTATION`	=	up-right
`TEXT_READING_DIRECTION`	=	left-right

and ground truth text strings that contain no UW-II escape sequences. The text in these zones, which we will call the *simple text-body* zones, consists solely of 92 standard ascii characters (space, A–Z, a–z, 0–9, left and right double quotes, and the 27 punctuation symbols `` !"#$%&`'()+,-./[]_:;<>?@{}= ``) and no subscripts, superscripts, overlines, underlines or dot sequences. The database contains 3400 simple text-body zones on 540 pages.

On the basis of previous experience with the UW-II database two of the 30 articles were excluded from the experiments. Article "6" (i.e. W06 and W16) seems to have been scanned from a bound journal and a significant amount of text at the spine-side page edges appears as unreadable black blotches in the scanned images. When these pages are decoded, the blotches typically result in a very large number of 'i' and 'l' insertion errors. The resulting large error rate per se is not grounds for rejection. However, the number of insertions varies, unpredictably, with the detailed appearance of the 'i' and 'l' templates. The variation is large enough that it masks correlation between recognition accuracy and template parameters that is apparent in the rest of the data.

Article "A" (i.e. W0A and W1A) is printed in a sans-serif font in which 'l' and 'I' are essentially indistinguishable. Thus, a decoder that is based solely on character shape (i.e. no lexicon or other constraints on the sequence of recognized characters) will make a large number of substitution errors. As with article "6", it is not the large error rate that is troublesome, but rather the unpredictable variation of error rate with template parameters.

The basic scenario in document-specific decoding is to train a set of templates using a subset of the pages to be recognized plus their transcriptions and then use the templates to decode the remaining pages. This scenario was enacted separately for each of the 28 selected UW-II articles in the following way.

The simple text-body zones for each article were divided into training and test sets by partitioning the article page images into two collections, each containing a complete set of article pages with approximately equal numbers of first and second generation copies in each set. This design insured that the training data for each article included all characters and fonts found in the test data and had similar image quality, while preserving the disjointness of the training and test sets. The total size of the training or test data for the 28 articles was about 606,000 glyphs if space characters are counted, or 519,000 glyphs if space is excluded.

Templates and channel parameters were estimated from the training zones using the supplied ground truth text as the training transcriptions. The test zones were then decoded using the estimated templates. The decoder was a simple horizontal text line decoder that placed no constraints on the sequence of recognized characters [3]. Text line baselines were estimated prior to decoding using a least-squares procedure, similar to that described in [4], to fit fifth order polynomials to glyph image positions obtained from a modified version of the commercial omni-font OCR program *TextBridge*.[1]

Table 1 summarizes character recognition performance using templates with 1–4 foreground levels. For comparison, results using *TextBridge* are also included. The columns labeled "#1" and "#2" identify the two most frequent substitution (e.g. 1 $\Rightarrow$ I), insertion (e.g. $\epsilon \Rightarrow sp$) or deletion (e.g. $sp \Rightarrow \epsilon$) errors and their frequencies.

Table 1(a) indicates that overall decoding accuracy using a single write-black foreground level (.752%) is comparable to the accuracy of *TextBridge* (.738%). The overall decoder error rate drops by almost a factor of four (to about .20%) when three or four foreground levels are used. For every number of foreground levels, insertion or deletion of space is one of

[1] Xerox Desktop Document Systems, Peabody, MA.

Foreground Levels (Mean)				Character Errors						
α_1	α_2	α_3	α_4	Subs.	Dels.	Ins.	Total	% Error	#1	#2
.871	—	—	—	3131	208	1217	4556	0.752	l $\Rightarrow$ I (974)	$\epsilon \Rightarrow sp$ (610)
.871	.000056	—	—	1597	242	378	2217	0.366	l $\Rightarrow$ I (649)	$sp \Rightarrow \epsilon$ (184)
.974	—	.440	—	1391	187	932	2510	0.414	$\epsilon \Rightarrow sp$ (577)	l $\Rightarrow$ I (216)
.974	.000056	.438	—	678	227	310	1215	0.201	$sp \Rightarrow \epsilon$ (178)	$\epsilon \Rightarrow sp$ (129)
.994	.000056	.341	.823	693	218	292	1203	0.199	$sp \Rightarrow \epsilon$ (170)	$\epsilon \Rightarrow sp$ (118)
TextBridge				2616	920	934	4470	0.738	$\epsilon \Rightarrow$ i (318)	u $\Rightarrow$ t (175)

(a)

Foreground Levels (Mean)				Character Errors						
α_1	α_2	α_3	α_4	Subs.	Dels.	Ins.	Total	% Error	#1	#2
.871	—	—	—	3077	58	606	3741	0.721	l $\Rightarrow$ I (973)	l $\Rightarrow$ 1 (299)
.871	.000056	—	—	1579	58	261	1898	0.366	l $\Rightarrow$ I (649)	i $\Rightarrow$ l (142)
.974	—	.440	—	1384	42	310	1736	0.335	l $\Rightarrow$ I (216)	l $\Rightarrow$ 1 (181)
.974	.000056	.438	—	669	51	177	897	0.173	l $\Rightarrow$ I (68)	, $\Rightarrow$. (56)
.994	.000056	.341	.823	685	50	169	904	0.174	l $\Rightarrow$ I (107)	$\epsilon \Rightarrow$ n (58)
TextBridge				2588	777	817	4182	0.806	$\epsilon \Rightarrow$ i (312)	u $\Rightarrow$ t (179)

(b)

Table 1: Character recognition results for Univ. of Washington English Document Database II (simple text-only zones from 28 articles). (a) All characters. (b) Excluding space characters.

the two most frequent decoding errors. Moreover, comparison of tables 1(a) and 1(b) reveals that the space error rate is greater than the mean character error rate. Since the width of an interword space can vary greatly within a document, it is not surprising that using a rigid space template results in a significant number of errors. However, the error rate is not excessive, suggesting that treating space as just another character is reasonable.

The single most frequent decoding error for printing characters is misrecognizing 'l' as 'I'. This not surprising, given the similarity of 'l' and 'I' in many fonts and the absence of a language model in the decoder. It is believed that such errors could be reduced by using a lexicon or character n-gram model.

7. SUMMARY

A multilevel imaging model for DID has been described in which the pixels of the ideal image take values from a finite set of levels or colors. Multilevel versions of previous bilevel image sources, channels, decoders and template estimation procedures are developed. The potential advantage of multilevel templates for decoding is demonstrated through a factor of four reduction (compared with bilevel templates) in character error rate using the UW-II English journal database.

ACKNOWLEDGEMENTS

The author wishes to thank Phil Chou and Les Niles for many stimulating discussions throughout the course of this work. Ben Wittner of Xerox Desktop Document Systems provided the version of *TextBridge* used in the experiments.

References

[1] H. Baird, "Document image defect models", in *Structured Document Image Analysis*, H. S. Baird, H. Bunke, and K. Yamamoto, eds., pp. 546–556, Springer-Verlag, 1992.

[2] D. Bloomberg and L. Vincent, "Blur hit-miss transform and its use in document image pattern detection", in *Document Recognition II*, L. Vincent and H. Baird, editors, Proc. SPIE vol. 2422, pp. 84–97, 1995.

[3] A. Kam and G. Kopec, "Separable source models for document image decoding", in *Document Recognition II*, L. Vincent and H. Baird, editors, Proc. SPIE vol. 2422, pp. 84–97, 1995.

[4] G. Kopec, "Least-squares font metric estimation from images", *IEEE Trans. on Image Processing*, vol. 2, no. 4, Oct. 1993.

[5] G. Kopec and P. Chou, "Document image decoding using Markov source models", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, June, 1994, pp. 602–617.

[6] G. Kopec and M. Lomelin, "Document-specific character template estimation", in *Document Recognition III*, L. Vincent and J. Hull, editors, Proc. SPIE vol. 2660, pp. 14–26, 1996.

[7] G. Kopec, "Document image decoding in the UC Berkeley Digital Library, in *Document Recognition III*, L. Vincent and J. Hull, editors, Proc. SPIE vol. 2660, pp. 2–13, 1996.

[8] G. Nagy and Y. Xu, "Priming the recognizer", *Proc. Workshop on Document Analysis Systems (DAS-96)*, Malvern, PA, pp. 263–281, 1996.

[9] I. Phillips, S. Chen, J. Ha and R. Haralick, *Reference Manual for the UW English/Japanese Document Image Database II*, Version 2.01, ISL report, Dept. of Electrical Eng., Univ. of Washington, Seattle, WA, March 8, 1995.