

Bayesian Super-Resolution of Text in Video with a Text-Specific Bimodal Prior

Katherine Donaldson, Gregory K. Myers
SRI International

{katherine.donaldson; gregory.myers}@sri.com

Abstract

To increase the range of sizes of video scene text recognizable by optical character recognition (OCR), we developed a Bayesian super-resolution algorithm that uses a text-specific bimodal prior. We evaluated the effectiveness of the bimodal prior, compared with and in conjunction with a piecewise smoothness prior, visually and by measuring the accuracy of the OCR results on the variously super-resolved images. The bimodal prior improved the readability of 4- to 7-pixel-high scene text significantly better than bicubic interpolation, and increased the accuracy of OCR results better than the piecewise smoothness prior.

1. Introduction

A capability to automatically identify and extract the contents of video imagery would enable videos to be indexed in a convenient and meaningful way for later reference, and would enable actions (such as automatic notification and dissemination) to be triggered in real time by the contents of streaming video. Video text recognition, or video OCR, is a useful tool for characterizing the contents of video containing overlay text (text captions superimposed over the video imagery, such as in broadcast news programs) and scene text (text that appears in the real scene of the video, such as text on street signs, nameplates, and billboards). This paper focuses on the recognition of scene text, which is often too small or of too poor quality to be recognized by an OCR process. In particular, text fonts with capital letters that are less than eight pixels high do not have sufficient resolution to be reliably recognized by commercial OCR products. Even custom-developed character recognizers that have been specially designed for recognizing small overlay text do not perform well on text fonts with capital letters smaller than seven pixels [1,2].

Because text typically appears in the video scene longer than the duration of a single video frame, multiple image samples representing the same text can be captured and made available for processing.

Therefore, super-resolution techniques can be considered for producing an enhanced representation of the text image with the goal of extending the range of text sizes and qualities of scene text that can be recognized in video imagery. Super-resolution is a process in which a single high-resolution image is constructed from a set of lower-resolution images, which could be degraded and aliased. If there is some relative motion between the camera and the scene (e.g., the text is on a moving object, or the pose of the camera is changing), each of the low-resolution images represents a different sampling of the scene. Super-resolution works by combining the complementary information about the scene contained in each of the samplings.

There has been a substantial amount of previous work in super-resolution for general imagery. Chaudhuri [3] and Park et al. [4] provide good reviews of various super-resolution approaches. However, in our application the super-resolution processing can be restricted to specific regions of the image that are likely to contain text. Processes that extract and read text from imagery typically apply a text detection process first, so that subsequent OCR processing can be focused only on the regions of the imagery containing text. The performance of scene text detectors is becoming increasingly reliable [5]. Therefore, we are interested in investigating super-resolution approaches that use the knowledge that the region to be enhanced contains text.

Super-resolution has been applied to text imagery in the past, but in most cases the algorithms have not been tailored specifically for text. Li and Doermann [6] used the method of projection onto convex sets (POCS), which was based on work by Patti et al. [7], to deblur scene text. Capel and Zisserman [8] tested four estimators for the super-resolution enhancement of text: a maximum likelihood (ML) estimator; the iterative back-projection method of Irani and Peleg [9]; a maximum a posteriori (MAP) estimator that incorporated a piecewise smoothness prior with a Huber penalty function (originally developed by Schultz and Stevenson [10]); and an estimator regularized using the Total Variation norm [11]. The latter two estimators generated superior results

compared with those of the former two estimators. In later work, Capel and Zisserman [12] compared the performance of a maximum likelihood estimator with hard constraints on individual pixel values with that of a MAP estimator with smoothness constraints; the former yielded sharper results because of the lack of any imposed spatial correlation.

Another approach within a Bayesian framework uses example-based priors. These algorithms are trained on a specific class of images, such as text; the correspondence between low-resolution and high-resolution imagery is learned from samples in a training set, and it is assumed that the relationship is the same as or similar to that in the test set. Baker and Kanade [13] achieved good results by applying this algorithm to text images with the same fonts as those in the training images.

Chiang and Boulton [14] developed an alternative approach, an edge-based super-resolution technique that attempts to locate the edges to subpixel accuracy in a sequence of images taken of a scene with text, and then fuses the conglomerated edge information into the first image. By focusing on derivative information rather than 0th-order pixel values, they avoid any potential illumination problems. In another approach, Chiang and Boulton [15] used a noniterative algorithm that resamples and warps the images to create a set of aligned, upsampled images, which are then straight-forwardly fused; the fused result is then deblurred. This approach was applied to video imagery of text, and performance was measured in terms of character recognition rate.

Perhaps the most salient property of text is that it is generally bimodal. By its very nature, text characters must have some contrast with the background to make them human-readable. Therefore, for most text in real scenes, the intensities of the text pixels tend to cluster around one value, and the intensities of the background pixels tend to cluster around another (of course, there are exceptions, due to large illumination variations within the text region and highly stylized graphic and color design). Therefore, the algorithm described in this paper applies a bimodal prior within a MAP Bayesian super-resolution framework. The MAP Bayesian framework allows this a priori knowledge about the imagery to be introduced explicitly as constraints. This approach is similar to that of Cheeseman et al. [16]; Schultz and Stevenson [10]; Hardie et al. [17]; and Capel and Zisserman [8,12], except that a bimodal prior is applied instead of a smoothness prior. A bimodal constraint has been successfully applied to the resolution enhancement of text in single images by Thouin and Chang [18], who used a nonlinear optimization technique to maximize

the bimodal-smoothness-average score of the expanded image.

In this paper the performance of this approach is compared with that of a piecewise smoothness constraint as a prior in the computation, and no prior ML. Because our goal is better OCR performance and not merely a better-looking image, the results are quantitatively evaluated by running an OCR engine on the super-resolved result in addition to visually showing the results of super-resolution on text images. To assess performance under realistic imaging conditions, the algorithms are evaluated with video imagery taken with a common consumer-grade camera.

2. Algorithm Description

Our super-resolution algorithm is based on the basic Bayesian framework. Assuming a set of N low-resolution observation images $\mathcal{L} = \{L_k\}$, the algorithm finds the high-resolution image H such that the conditional probability of H , given the observed \mathcal{L} , $P[H | \mathcal{L}]$, is maximized. This is difficult to calculate directly, but using our camera model each L_k can be expressed in terms of H , which allows us to calculate $P[\mathcal{L} | H]$. Using Bayes' law, we obtain $P[H | \mathcal{L}] = P[\mathcal{L} | H] \cdot P[H] / P[\mathcal{L}]$. The denominator, $P[\mathcal{L}]$, does not affect the maximization. Therefore, to find the most probable high-resolution image, given the observation images, we need to find the high-resolution image H that maximizes $P[\mathcal{L} | H] \cdot P[H]$, which is the MAP estimator.

If it is assumed that, barring any observed low-resolution images, all high-resolution images are equally likely, this algorithm could be further simplified to finding H that maximizes $P[\mathcal{L} | H]$, which is the ML estimator. But we have assumed that the set of high-resolution images contains only images of binary text scenes, so priors for bimodality, $P_B[H]$, and piecewise smoothness, $P_S[H]$ can be derived.

Once $P[\mathcal{L} | H]$ has been defined, the super-resolved image, H , that maximizes $P[H | \mathcal{L}]$, is iteratively calculated by stepping down the gradient of the negative log likelihood of $P[\mathcal{L} | H] \cdot P[H]$ until a minimum is reached or a maximum number of iterations are executed.

The final step in the algorithm is to dynamically threshold the image so that the binary result can be fed to the OCR engine. The algorithm chooses the dynamic threshold by fitting Gaussians to the foreground (black) and background (white) pixel distributions using the expectation maximization (EM)

algorithm [19] and calculating the value midway between their means.

2.1. Modeling and Image Formation

As we focus on images of text, we assume that the text is on a plane. We are currently using a simplified 2-D version of our camera model to test the effectiveness of our priors. This simplified model assumes that the scene text plane is perpendicular to the camera and that all motion is translational. The model can be directly extended to a 3-D projective transform at the cost of additional processing time and added complexity due to increased degrees of freedom in the registration and projection steps; however, we are focusing on the relative advantages of our priors, so we use the simplified camera model.

Since it is assumed the text plane is perpendicular to the camera with only translational motion, the image formation process can be represented as a translation of H , the high-resolution image, with respect to the camera; then a convolution with a pillbox blur kernel representing the camera aperture, followed by summing over the active areas of the camera CCD pixels to account for subsampling and fill factors. Illumination variations are handled with an optional pre-processing step to remove x and y linear illumination gradients by using a least squares fit on the image values. The image formation process can thus be written

$$L_k = \Gamma(\delta \otimes H(x - x_k, y - y_k)) \quad , \quad (1)$$

where L_k is the k^{th} observed low-resolution image, Γ is the subsampling lattice, δ is the pillbox blur kernel, (x_k, y_k) is the translation vector for L_k , and H is the high-resolution image.

If H and L_k are reordered into one-dimensional vectors, using a lexicographic ordering, the transformation between them can be represented as the sparse correspondence matrix A_k , and the image formation process can be written

$$L_k = A_k \bullet H \quad . \quad (2)$$

In practice, the computation of the projection matrices A_k is the most costly step of the algorithm, although once they are calculated they are used repeatedly to determine the error of the estimated H in the gradient descent iterations. We have chosen to compute A_k columnwise, where every column of A_k is the mapping of one pixel of H into L_k . To simplify the computation, the projection is calculated by using a numerical integration over a fine grid on the low-resolution camera plane. The default integration grid resolution is set to be a factor of 5 greater than the resolution increase from L_k to H . Assuming a

resolution increase of $2\times$ to $5\times$, typical integration grid resolutions range from 10 to 25 cells across each low-resolution camera pixel. To fill in one column of the matrix A_k , we need to calculate what fraction of the corresponding high-resolution pixel maps to each low-resolution pixel. Thus, the projection onto the camera plane grid of each high-resolution pixel in H is calculated; with our simplified camera model this projected pixel is a translated square. To fill in the projected square on the integration grid, each grid cell is set to the fraction by which it overlaps with the projected high-resolution pixel. The projection is then blurred by convolving it with the pillbox blur kernel δ , which has been precomputed on the same fine resolution integration grid. Typically, the camera blur is on the order of one low-resolution pixel, which makes its gridded representation 10 to 25 cells wide. The integration of this fine grid over the active area of each low-resolution pixel yields the values for a column of A_k . As an optimization, since we are using a translational camera model we can save the projection of the first pixel of H onto the integration grid, translate every other pixel in H the appropriate amount, and integrate it over the active area of the low-resolution pixels.

2.2. Registration

Taking advantage of the simplified camera model, we used pairwise correlation with quadratic interpolation and least-squares fit to determine the subpixel translation vector (x_k, y_k) for each observed low-resolution image. Each of the N low-resolution images, L_k , in the set \mathcal{L} is correlated against every other image in the set \mathcal{L} , yielding $(N^2 - N)/2$ correlation matrices. A quadratic interpolation around the low-resolution pixel peak is then used to find the subpixel peak of each correlation matrix, yielding a set of observed subpixel translation vectors $\{m_{i,j}\}$, where $m_{i,j}$ is the translation from L_i to L_j . The information from all of the pair-wise correlations is combined, using a least-squares fit. Since translational motion is additive, the translation (x_{ik}, y_{ik}) between images L_i and L_k is equal to the translation from L_i to L_j (x_{ij}, y_{ij}) , plus the translation from L_j to L_k , (x_{jk}, y_{jk}) . Thus, using the independently measured translations $\{m_{i,j}\}$ we can construct a set of $(N^2 - N)/2$ linear equations of the form $\epsilon_{j,i} = (x_{0,j}, y_{0,j}) + m_{j,i} - (x_{0,i}, y_{0,i})$, where $m_{j,i}$ is the measured correlation translation between L_j and L_i , and $\epsilon_{j,i}$ is the error in the estimation, and $(x_{0,i}, y_{0,i})$ are the N unknowns. Given these equations a standard least-squares error minimization is used to estimate the values for $(x_{0,i}, y_{0,i})$ that minimize the error values $\epsilon_{j,i}$.

The registration accuracy of this method was measured at roughly ± 0.05 to ± 0.1 pixel on simulated data.

2.3. ML Estimator

Making the assumption that all H , prior to observing \mathcal{L} , are equally likely, the ML estimator calculates the super-resolved image solely from the likelihood of the observed low-resolution images, given an estimate of the high-resolution image, $P[\mathcal{L}|H]$. Assuming that the image noise is Gaussian with a mean of zero and variance σ_n^2 , the probability of observing the low-resolution image L_k , given an estimate of the high-resolution image H' , is

$$P[L_k|H'] = \prod_{i,j} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(L_{i,j,k} - L'_{i,j,k})^2}{2\sigma_n^2}\right) \cdot \quad (3)$$

This is a product of normal distributions with noise variance σ_n^2 , with the deviation from the mean being the estimate error $(L_{i,j,k} - L'_{i,j,k})$, over image coordinates (i,j) . L'_k is an estimate of L_k , calculated by projecting the current estimate of the high-resolution image H' , using the previously defined correspondence matrix A_k from Eq. (2).

The maximum likelihood estimate H_{ML} is calculated by minimizing the negative log likelihood over all observed images \mathcal{L} .

$$\begin{aligned} H_{ML} &= \operatorname{argmax}_{H \in \mathcal{H}} \prod_k P[L_k|H'] \\ &= \operatorname{argmin}_{H \in \mathcal{H}} \sum_k -\log(P[L_k|H']) \\ &= \operatorname{argmin}_{H \in \mathcal{H}} \sum_{i,j,k} (L_{i,j,k} - L'_{i,j,k})^2 / (2\sigma_n^2) \\ &= \operatorname{argmin}_{H \in \mathcal{H}} \sum_{i,j,k} (L_{i,j,k} - L'_{i,j,k})^2 \cdot \quad (4) \end{aligned}$$

2.4. MAP Estimators

Forming a MAP estimator requires stating the a priori expectations of H as a prior probability distribution. From the two features of binary text images, bimodality and piecewise smoothness, two prior probability distributions on the high-resolution image, $P_B[H]$ and $P_S[H]$, are derived.

2.4.1. Bimodality Prior. For the bimodality prior we used an exponentiated fourth-order polynomial with maxima at the calculated centers of the black and white pixel distributions. We use the following exponential-based bimodal distribution, seen in Fig. 1 and Eq. (5), so that it can easily be simplified when the log of $P_B[H'_{i,j}]$ is calculated.

$$P_B[H'_{i,j}] = c_b(\sigma_b, \mu_0, \mu_1) \cdot \exp\left(\frac{(H'_{i,j} - \mu_0)^2 \cdot (H'_{i,j} - \mu_1)^2}{\sigma_b^4}\right), \quad (5)$$

where μ_0 and μ_1 are current estimates of the positions of the two peaks of the bimodal distribution, σ_b determines the width of the black and white distributions, and $c_b(\sigma_b, \mu_0, \mu_1)$ is a normalizing constant for the distribution dependent on σ_b , μ_0 , and μ_1 . We automatically determine μ_0 and μ_1 by using an expectation maximization algorithm [19] to fit two Gaussians to the histogram of the current super-resolution estimate. An estimate of μ_0 and μ_1 is computed between each gradient descent step. This prior function only somewhat approximates the actual distribution of pixel values in an image region containing scene text. In particular, it exactly models the distribution only when the proportions of black and white pixels are equal, and their variances are also equal. However, this functional form makes it quite easy to compute the log likelihood function, which we deemed a reasonable trade-off.

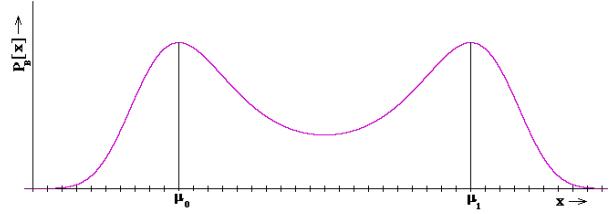


Figure 1. Bimodality prior $P_B[x]$ vs. image intensity.

2.4.2. Smoothness Prior. Since most text images are locally smooth with step discontinuities, for a smoothness prior we use a Gibbs prior with a Huber gradient penalty function, similar to that described by Schultz and Stevenson [10] and Hardie et al. [17]. The Gibbs prior represents piecewise smooth data with the probability density defined as

$$P_S[H'_{i,j}] = c_s \cdot \exp\left(\rho(H'_{i,j} - \bar{H}'_{i,j}) / \sigma_s^2\right), \quad (6)$$

where c_s is a normalizing constant, $\rho(x)$ is the Huber edge penalty function, and $H'_{i,j} - \bar{H}'_{i,j}$ is a local measure of image smoothness (which is small where the image is smooth, and large where it is discontinuous). $\bar{H}'_{i,j}$ is the average of the four nearest neighbors of $H'_{i,j}$:

$$\bar{H}'_{i,j} = (H'_{i-1,j} + H'_{i+1,j} + H'_{i,j-1} + H'_{i,j+1}) / 4 \cdot \quad (7)$$

The likelihood of discontinuities in the data is controlled by the Huber edge penalty function $\rho(x)$ (see Fig. 2):

$$\rho(x) = \begin{cases} x^2, & |x| \leq \alpha \\ 2\alpha|x| - \alpha^2, & |x| > \alpha; \end{cases}, \quad (8)$$

where α is the threshold between the quadratic and linear regions. This makes larger discontinuities much more likely than they are with a strictly quadratic edge penalty. In the linear region the derivative of this penalty function is constant, which thereby preserves the steep edges in the image when the algorithm steps along the gradient of the log likelihood.

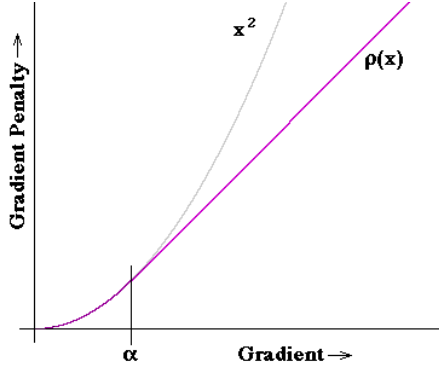


Figure 2. Penalty function.

2.4.3. Three MAP Priors. Replacing the uniform $P[H]$ of the ML estimator with the bimodality and smoothness priors leads to three different maximum a posteriori estimators. The first uses the bimodal prior $P_B[H'_{i,j}]$ only:

$$H_{MAP-B} = \operatorname{argmax}_{H \in \mathcal{H}} \prod_k P[L_k | H'] \cdot \prod_{i,j} P_B[H'_{i,j}]. \quad (9)$$

The second uses the smoothness prior $P_S[H'_{i,j}]$ only:

$$H_{MAP-S} = \operatorname{argmax}_{H \in \mathcal{H}} \prod_k P[L_k | H'] \cdot \prod_{i,j} P_S[H'_{i,j}]. \quad (10)$$

The third uses both the smoothness prior and the bimodal prior:

$$H_{MAP-SB} = \operatorname{argmax}_{H \in \mathcal{H}} \prod_k P[L_k | H'] \cdot \prod_{i,j} P_B[H'_{i,j}] \cdot P_S[H'_{i,j}]. \quad (11)$$

The maximum a posteriori estimates, H_{MAP-B} , H_{MAP-S} , and H_{MAP-SB} , are calculated by minimizing the negative log likelihood of H over all observed images, as was done with the ML estimator. Here only H_{MAP-SB}

is shown, because it is a superset of the other two MAP estimators. Expanding H_{MAP-SB} and taking the negative log yields

$$H_{MAP-SB} = \operatorname{argmin}_{H \in \mathcal{H}} \sum_{i,j,k} \left[(L_{i,j,k} - L'_{i,j,k})^2 / (2\sigma_n^2) \right] + \sum_{i,j} \left[(H'_{i,j} - \mu_0)^2 \cdot (H'_{i,j} - \mu_1)^2 / \sigma_b^4 - \rho(H'_{i,j} - \bar{H}'_{i,j}) / \sigma_s^2 \right], \quad (12)$$

where σ_n , σ_b , and σ_s have effectively become the weights between the error term and the two priors.

3. Results

3.1. Experimental Procedure

All test images were taken with Sony miniDV cameras with image stabilization disabled. These cameras use the compressed DV25 format to record video data. Digital video (DV) compression uses discrete cosine transform to compress pixel data by a factor of 5:1 (in the DV25 format) on an intraframe basis. DV compression is relatively light but is not lossless; it produces ringing artifacts around high-contrast edges such as those in text images, and other artifacts. To be useful in practical applications, however, the super-resolution algorithm must tolerate these artifacts gracefully, since DV25 is the native format of almost all consumer-grade digital video cameras. We also show some results on data that were further compressed by MPEG-1 software.

The OCR engine used to evaluate the super-resolution results was the MTX engine within Scansoft's DevKit 2000.

The algorithm parameters were experimentally determined. The fill factor for the video camera's CCD model was set at a reasonable default of $60\% \times 60\%$ for all tests. The blur diameter of the camera model and the weighting of the bimodal and smoothness priors were set by hand for each video sequence. With nominally chosen values for σ_n , σ_b , and σ_s , the algorithm was run with blur diameters from 0.5 to 1.5 camera pixels in steps of 0.25 pixels, and the most visually acceptable blur was chosen. Once the blur was chosen, the code was run over a range of σ_n , σ_b , and σ_s values, and the image with the best subjective noise vs. detail trade-off was chosen.

This hand tuning of parameters is clearly a suboptimal solution, and will not be an option if the entire process were to be completely automated. The

blur parameter could be estimated using a blind image deconvolution, preferably an algorithm that takes into account the underlying binary nature of the scenes. Another possibility would be to incorporate the estimation into the super-resolution process [8], although this would require the re-estimation of the projection matrices A_k (which is very time consuming). With the σ parameters the eventual goal is to use a diverse set of images to experimentally determine the relationship between σ_n , σ_b , and σ_s , and the OCR accuracy. Possibly the optimal settings for σ_n , σ_b , and σ_s will depend on the initial text height, the resolution increase factor between L and H , and/or the estimated camera blur.

In addition to being lightly compressed, DV25 is also an interlaced video format. Accordingly, the fields were separated and considered to be two independent images. To compensate for these double images, the resolution increase factor in the y dimension was doubled, and the camera CCD fill factor to $60\% \times 30\%$ was decreased. This is not necessary with the MPEG-1 compressed video, which has already been de-interlaced and subsampled as part of the compression process.

3.2. Pixel Height vs. Readability, Using MAP-B

Fig. 3 shows a full frame extracted from a video showing text of various sizes from the Gettysburg Address. Table 1 shows the results of a test in which the super-resolution process uses up to 32 input images of various-size text. Fig. 4 shows OCR accuracy (in terms of the fraction of characters correctly recognized) as a function of text height for various numbers of low-resolution frames, with an additional curve showing the OCR accuracy of the images after expansion via bicubic interpolation.

Under low blur, noise, and distortion conditions, the benefits of the MAP estimator with the bimodal prior become apparent when the pixel heights of the capital letters of the observed text are between 4.5 and 6.75 pixels. The benefits disappear at letter heights of about 9 pixels, since at that height the OCR system becomes able to recognize the text perfectly, using only bicubic interpolation and binarization as preprocessing.

When the Gettysburg Address video sequence is compressed by an MPEG-1 scheme, the resolution is halved to 320×240 . The 5.25-pixel-high text is thus reduced to a height of 2.6 pixels. Fig. 5 demonstrates the effectiveness of the algorithm on the compressed video data. The character recognition rate improved significantly for text approximately 3 to 7 pixels high.

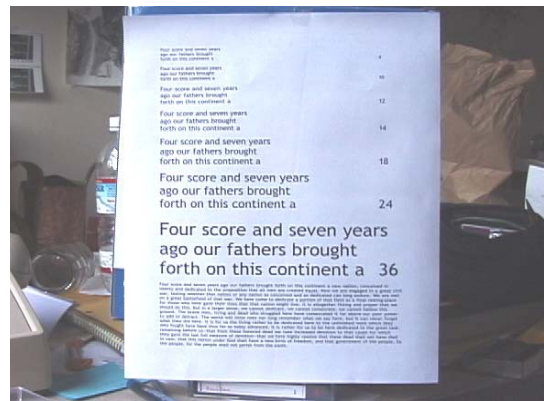


Figure 3. Full video frame of text.

Table 1. Super-resolution output of the MAP-B estimator.

Text Height (pixels)	Original Frames	Images Super-resolved via MAP-B Estimator
3.75		
4.5		
5.25		
6.75		
9		

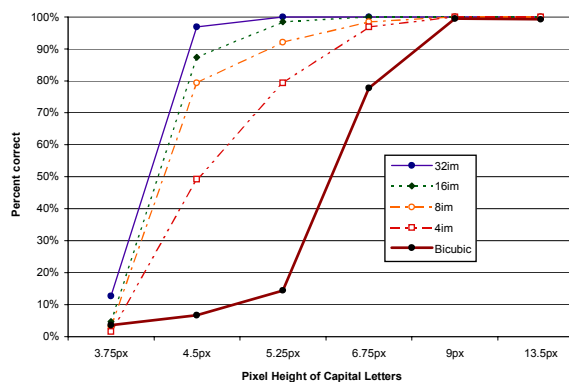


Figure 4. OCR accuracy vs. text height achieved with the MAP-B estimator.

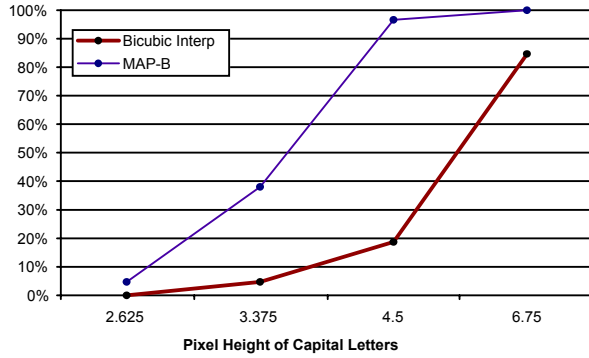


Figure 5. OCR accuracy vs. text height of MPEG-1-compressed imagery.

3.3. Comparison of Priors' Effect on Recognition

3.3.1. Gettysburg Text Image. Table 2 shows the recognition results of the two smallest texts in the Gettysburg Address video frame (3.75 and 4.5 pixels) for the different priors. The estimators were run on 32 images, with $\sigma_n = 1$, $\sigma_b = 12$, $\sigma_s = 12$, and blur diameter = 1 pixel. The recognition improves somewhat when the smoothness prior is added to the ML estimator, and improves more when the bimodal prior is added. In these cases it can be seen that the advantage of the smoothness prior is subsumed in the advantage of the bimodal prior; if the bimodal prior is being used, adding the smoothness prior does not yield any additional accuracy in the OCR.

Table 2. Recognition results for various estimators.

Estimator	Character Height in Pixels	
	3.75	4.5
ML	25%	91%
MAP-S	27%	94%
MAP-B	34%	98%
MAP-BS	34%	98%

3.3.2. Text on License Plates. The video used in this test is closer to a typical sequence in surveillance videos than those discussed above. It consists of a panning shot across a parking lot, showing text on license plates. Fig. 6 shows one frame from this video sequence, and Fig. 7 shows the results for the various priors. The estimators were run on 16 images, with $\sigma_n = 1$, $\sigma_b = 6$, $\sigma_s = 1$, and blur diameter = 0.75 pixels.



Figure 6. Full video frame.

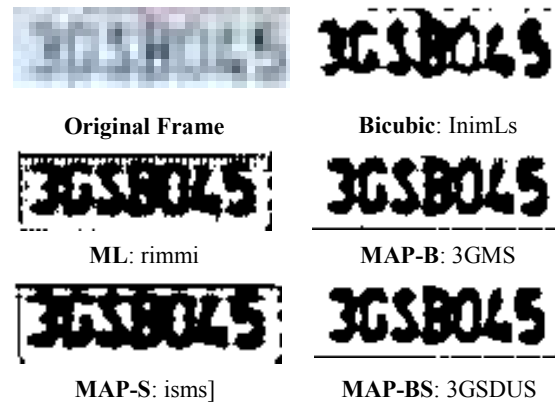


Figure 7. Results for various estimators

3.3.3. Discussion. In general, the largest improvements in OCR performance seem to result from increases in character separation rather than from the reduction of general noise. In some cases the super-resolved image can be improved cosmetically, its edges sharpened, and the speckle noise reduced without improving the OCR performance. As the bimodal prior has a larger effect on character separation than the piecewise smoothness prior, it is more effective at improving the OCR performance. The piecewise smoothness prior is good at reducing false speckle details in the results and halting the gradient descent before it reaches the regime of noise amplification instead of detail amplification (unlike the ML estimator only), but it does not appear to be especially good at separating closely spaced characters. Due to the method the piecewise smoothness prior uses to preserve sharp transitions, it does not move the location of these transitions and therefore has little effect on the thresholded images. In contrast, when the bimodal prior raises the values of some pixels near the gray edges of transitions toward black, it causes the error constraint to push other nearby pixels toward white in order to balance the average to the observed low-resolution gray value. In this respect, the estimators

that use the bimodality prior (MAP-B and MAP-BS) seem to have a greater impact on performance.

4. Summary

To increase the range of sizes of video scene text processed by OCR, we developed a Bayesian super-resolution algorithm that uses a text-specific bimodal prior. We quantitatively evaluated the results by running an OCR engine on the super-resolved result. We tested two prior distributions on high-resolution imagery, a bimodal prior, and a smoothness prior. The MAP estimator using the bimodality prior significantly increases the accuracy of the OCR results (compared with bicubic interpolation) on scene text 4 to 7 pixels high, when processing DV compressed data, and on text between 3 and 7 pixels high, when processing MPEG-1 compressed data. To our knowledge, the recognition of such small text has not been previously reported. Our tests also show that the bimodality prior is more effective in increasing the accuracy of OCR results than the piecewise smoothness prior.

5. Acknowledgement

This material is based on work supported in whole by the Advanced Research and Development Activity (ARDA). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

6. References

[1] Aradhye, H., C. Dorai, and J. Shim, "Study of Embedded Font Context and Kernel Space Methods for Improved Videotext Recognition," in *IEEE Intl. Conf. Image Processing*, 2001.

[2] Dorai, C., H. Aradhye, and J. Shim, "End-to-End Videotext Recognition for Multimedia Content Analysis," in *IEEE Intl. Conf. Multimedia and Expo*, 2001.

[3] Chaudhuri, S., *Super-Resolution Imaging*, Kluwer International Series in Engineering and Computer Science, No. 632, ISBN 0-7923-7471-1, September 2001.

[4] Park, S.C., M.K. Park, and G.K. Moon, "Super-Resolution Image Reconstruction: A Technical Overview," *Signal Processing Magazine*, IEEE, Vol. 20, Issue 3, pp. 21–36, May 2003.

[5] Doermann, D., J. Liang, and H. Li, "Progress in Camera-Based Document Image Analysis," *Proc. Intl. Conf. on Document Analysis and Recognition*, pp. 606–16, 2003.

[6] Li, H., and D. Doermann, "Superresolution-Based Enhancement of Text in Digital Video," *Proc. Intl. Conf. on Pattern Recognition*, pp. 847–50, 2000.

[7] Patti, A., M. Sezan, and A. Tekalp, "Superresolution Video Reconstruction with Arbitrary Sampling Lattices and Nonzero Aperture Time," *IEEE Trans. on Image Processing*, Vol. 6, No. 8, pp. 1064–76, August, 1997.

[8] Capel, D., and A. Zisserman, "Super-Resolution Enhancement of Text Image Sequences," *Proc. Intl. Conf. Pattern Recognition*, Vol. 1, pp. 600–5, September 2000.

[9] Irani, M., and S. Peleg, "Improving Resolution by Image Restoration," *Computer Vision, Graphics, and Image Processing*, Vol. 53, pp. 231–9, 1991.

[10] Schultz, R., and R.L. Stevenson, "Extraction of High-Resolution Frames from Video Sequences," *IEEE Trans. Image Processing*, Vol. 5, No. 6, pp. 996–1011, June 1996.

[11] Vogel, C.R., and M.E. Oman, "Fast, Robust Total Variation-Based Reconstruction of Noisy, Blurred Images," *IEEE Trans. Image Processing*, Vol. 7, No. 7, pp. 813–24, July 1998.

[12] Capel, D., and A. Zisserman, "Super-Resolution from Multiple Views Using Learnt Image Models," in *Proc. CVPR*, 2001.

[13] Baker, S., and T. Kanade, "Limits on Super-Resolution and How to Break Them," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 24, No. 9, pp. 1167–83, September 2002.

[14] Chiang, M., and T.E. Boult, "Local Blur Estimation and Super-Resolution," *Proc. CVPR*, pp. 821–6, June, 1997.

[15] Chiang, M.-C., and T.E. Boult, "Efficient Super-Resolution Via Image Warping," *Image and Vision Computing*, Vol. 18, Issue 10, pp. 761–71, 2000.

[16] Cheeseman, P., B. Kanefsky, R. Knaft, J. Stutz, and R. Hanson, "Super-Resolved Surface Reconstruction from Multiple Images," *Maximum Entropy and Bayesian Methods*, ed. G. Heidbreder, Kluwer, pp. 293–308, 1996.

[17] Hardie, R.C., K.J. Barnard, and E.A. Armstrong, "Joint MAP Registration and High-Resolution Image Estimation Using a Sequence of Undersampled Images," *IEEE Trans. Image Processing*, Vol. 6, No. 12, pp. 1621–33, 1997.

[18] Thouin, P., and C.-I. Chang, "A Method for Restoration of Low-Resolution Text Images," *Proc. 1999 Symp. Document Image Understanding Technology*, Annapolis, Maryland, pp. 143–8, 1999.

[19] Dempster, A.P., Laird, N.M., and Rubin, D.B., "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Society, Series B*, Vol. 39, No. 1, pp. 1–38, 1977.