# Text Enhancement in Projected Imagery

Xiaodan Hu[1], Mohamed A. Naiel[1], Zohreh Azimifar[2], Ibrahim Ben Daya[1], Mark Lamm[3] and Paul Fieguth[1]
[1]University of Waterloo, Waterloo, Canada
[2]Shiraz University, Shiraz, Iran
[3]Christie Digital Systems Canada Inc., Kitchener, Canada
Email: {x226hu, mohamed.naiel, azimifar, ibendaya, pfieguth}@uwaterloo.ca, mark.lamm@christiedigital.com

## Abstract

There is great interest in improving the visual quality of projected imagery. In particular, for image enhancement, we would assert that text and non-text regions should be enhanced differently in seeking to maximize perceived quality, since the spatial and statistical characteristics of text and non-text images are quite distinct. In this paper, we present a text enhancement scheme based on a novel local dynamic range statistical thresholding. Given an input image, text-like regions are obtained on the basis of computing the local statistics of regions having a high dynamic range, allowing a pixel-wise classification into text-like or background classes. The actual enhancement is obtained via class-dependent Wiener filtering, with text-like regions sharpened more than the background. Experimental results on four challenging images show that the proposed scheme offers a better visual quality than projection without enhancement as well as a recent state-of-the-art enhancement method.

## 1 Introduction

Using a low-resolution projector to project high-resolution content is of great interest, since such an approach can significantly reduce the cost of projector display systems [1, 2]. In the literature, several methods have been proposed in order to achieve this goal [1–4]. Allen and Ulichney [2] used a low-resolution projection system with an optomechanical image shifter to reproduce an enhanced high-resolution content on a given screen and referred to this method as Wobulation, where the main idea is superimposing two low resolution images in rapid succession to reconstruct a higher resolution image. Later in 2015, Barshan *et al.* [5] proposed a resolution enhancement method, shifted superposition (SSPOS), by optimally learning a pair of spatially shifted low-resolution sub-images. Although the methods in [2, 5] increased the perceived image resolution, they did not consider the blurring effect caused by the projector-lens system. Recently, Ma *et al.* [6] proposed a resolution enhancement method by introducing a Wiener deconvolution filtering; for reasons of implementation simplicity, rather than a frequency-domain, i.e., 2D-DFT operation, the Wiener filter was approximated in the spatial-domain. In exploring possible Wiener filters, it was clear that effective filters for text tended to create artifacts in other content, and effective filters for imagery tended to under-enhance text.

Since text is so frequently embedded in displayed content, and furthermore, since text readability is essential in effectively conveying relevant information, the effective enhancement of text is of significant importance and represents significant added value for projector display systems. As a result, we have two basic problems at hand:

1. The detection of text-like regions (Section 3.1).
2. The enhancement of text-like regions (Section 3.2).

In this paper, a novel text-like region enhancement scheme for projector-based systems is introduced. We first propose a text-like detection method using local dynamic range statistical thresholding to generate a binary mask to segment text-like regions from the background. Then, two separate Wiener deconvolution filters are used to sharpen text-like regions and other regions, respectively. Applying more sharpening to text-like regions than on other content parts results in a better visual quality for the projected content. It is shown from the sample results in Figure 1 that unlike the method in [6], the proposed scheme is able to enhance the text-like regions as well as the background. Additional experimental results conducted on four challenging images show that the proposed method consistently provides a better quality than that offered by projection without enhancement as well as the recent state-of-the-art enhancement method in [6].

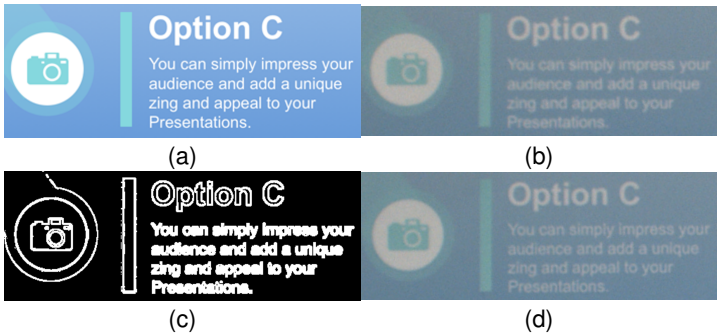


*Fig. 1:* Sample qualitative comparison on a test image (a), Wiener deconvolution method [6] (b), text mask of the proposed method (c), and the proposed text enhancement method (d).

## 2 System Model

The main idea of the proposed text enhancement scheme is to enhance text and background regions differently based on an identified text mask and two class-dependent Wiener deconvolution filters of different sharpening levels for text and non-text regions. As shown in Figure 2, the proposed scheme consists of three main parts: text detection, text and non-text regions filtering using Wiener deconvolution, and two-branch high-resolution superimposed projection. In order to achieve a two-position high-resolution projection for an input image $I$ of size $N_1 \times N_2$ using a low-resolution projector, we first up-sample the input image by a factor of $s$ in both $x$ and $y$ directions to obtain the up-sampled image $I_u$ of size $\hat{N}_1 \times \hat{N}_2$, where $\hat{N}_1 = \lfloor sN_1 \rfloor, \hat{N}_2 = \lfloor sN_2 \rfloor$ and $s = \sqrt{2}$ has been used. Next, the proposed text-like detection scheme, described in Section 3.1, is employed on $I_u$ in order to obtain the text mask, $M_T$, and thus, distinguish text-like regions from the background. Then, the enhanced image $\hat{I}_u$ is obtained by highly and moderately sharpening the up-sampled image $I_u$ using two different Wiener deconvolution kernels $\tilde{g}_T$ and $\tilde{g}_\Omega$ for the text and background components, respectively, as shown in Section 3.2.

As in [6], two sub-images $L_1$ and $L_2$ each of size $\tilde{N}_1 \times \tilde{N}_2$, which are needed for a low-resolution projector [5], are generated by first shifting $\hat{I}_u$ one pixel in both $x$ and $y$ directions, and then down-sampling $\hat{I}_u$ and its shifted version by a factor of $s^2$, where $\tilde{N}_1 = \lfloor \hat{N}_1/s^2 \rfloor$, and $\tilde{N}_2 = \lfloor \hat{N}_2/s^2 \rfloor$. Finally, similar to [5] the two sub-images are superimposed to project perceived high-resolution contents.

## 3 Methodology

In this section, the proposed text-like region detection method is introduced and a Wiener deconvolution-based text enhancement method is described. For text enhancement, a method based on statistics of dynamic range of local pixels and two class-dependant Wiener deconvolution filters is introduced.

### 3.1 Local Dynamic Range Statistical Thresholding

In the literature, text detection methods can be classified into two main categories. First, connected component-based methods have gained high attention in many studies [7–9]. For instance, [7] introduced the idea of Maximally Stable Extremal Regions (MSER), which partitions an input image into indivisible components, followed by classifying each component as text or background. However, it is known that MSER is sensitive to image blurring and is computationally expensive [10].

On the other hand, many text detection methods based on image thresholding were introduced [11–14]. For instance, in [12], a local horizontal differential filter and a thresholding scheme were performed in order to find vertical edges in an image. This method,

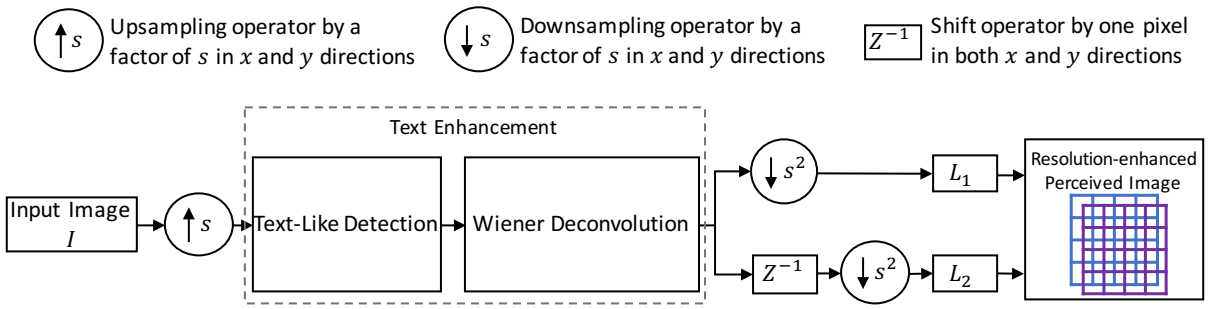


*Fig. 2:* An overview block diagram of the proposed text enhancement scheme.

however, lacks representing the 2D structures within the text regions. Later in [13], a simple binarization algorithm for extracting text regions was introduced. However, this method assumes that text regions are brighter than the background, which is not always true.

Text regions are always expected to be of higher contrast than their nearby background regions, since text is usually used to convey an important information. In this paper, a method for enhancing text-like regions by sharpening these regions more than that of the background ones is proposed. One way to approach this objective is to develop a robust text detection technique to localize a wide variety of text-like regions of several challenges such as different shapes, colors, fonts, styles, sizes, and orientations. In this section, a local dynamic range method is introduced to represent the contrast between the pixels of text and non-text regions, and use the statistics of local dynamic range to classify pixels into either text or non-text class.

In this method, the local maximum and minimum pixel values of the $i^{th}$ local sliding window $\aleph_i$ of size $k \times k$ are used to represent the level of contrast of a given input image $I_u$, where $i$ is the center of $\aleph_i$, $i = [(1,1),(1,2),...,(\hat{N}_1,\hat{N}_2)]$, $\hat{N}_1$ and $\hat{N}_2$ are the number of rows and columns of input image $I_u$. The local dynamic range $D(i)$ of the $i^{th}$ pixel in $I_u$ can now be obtained as:

$$D(i) = \max_{j \in \aleph_i}(I_u(i,j)) - \min_{j \in \aleph_i}(I_u(i,j)) \tag{1}$$

Let $h_d$ be the histogram of the dynamic range $D$ at every $d \in D$. In order to filter the background noise of low contrast, we compute the threshold $\tau_1$ as follows:

$$\tau_1 = \{d^* \in D : h_{d^*} < h_d, d^* \neq d, d \in D\} \tag{2}$$

where $h_{d^*}$ is the first strict local minimum of $h_d$. In the meantime, a 2D Gaussian filter of size $\hat{k} \times \hat{k}$ is used for reducing the effect of the outliers that may exist in the dynamic range matrix $D$ and the filtered dynamic range is denoted as $\hat{D}$. We now apply a thresholding operation to the smoothed dynamic range $\hat{D}$ as:

$$\tilde{D}(i) = \begin{cases} \hat{D}(i), & \hat{D}(i) > \tau_1 \\ \tau_1, & \text{Otherwise} \end{cases} \tag{3}$$

Then, the local statistics of the $i^{th}$ element of the thresholded dynamic range $\tilde{D}$ are obtained as:

$$\mu(i) = \frac{1}{k^2} \sum_{j \in \aleph_i} (\tilde{D}(j)) \tag{4}$$

$$\sigma(i) = \sqrt{\frac{1}{k^2} \sum_{j \in \aleph_i} (\tilde{D}(j) - \mu(i))^2} \tag{5}$$

where $\mu$ and $\sigma$ are the mean and the standard deviation of $\tilde{D}$. The final threshold $\tau_2$ for $\tilde{D}$ is obtained as:

$$\tau_2(i) = \mu(i) + \sigma(i) \tag{6}$$

The mask of text-like regions, $M_T$, is obtained by applying threshold $\tau_2$ on $\tilde{D}$ as:

$$M_T(i) = \begin{cases} 1, & \tilde{D}(i) > \tau_2 \\ 0, & \text{Otherwise} \end{cases} \tag{7}$$

### 3.2 Spatial-based Wiener Deconvolution Filtering

Wiener deconvolution filtering has been widely used to enhance the spatial resolution of images [15]. However, the filtering process works in the transform domain and involves domain transformations, which results in high computational complexity. Recently, Ma *et al.* [6] introduced a spatial kernel derived from the Wiener deconvolution filter and the filtering operation is simplified to be a 2D convolution in the spatial domain. In this section, we give a brief overview of the method in [6] and how we applied it in the proposed scheme. Given the estimated projector's point spread function (PSF) in the 2D-DFT domain denoted as $H(u,v)$, where $u$ and $v$ represent the frequency indices, and a certain signal-to-noise ratio (SNR), the Wiener deconvolution filter $G(u,v)$ is computed as:

$$G(u,v) = \frac{1}{H(u,v)} \left[ \frac{|H(u,v)|^2}{|H(u,v)|^2 + \frac{1}{SNR}} \right] \tag{8}$$

where $G$ and $H$ are of size $r \times r$. It is shown in [6] that obtaining the spatial kernel $g(n,m)$ by directly applying the inverse 2D-DFT on (8) causes over-sharpening artifacts for the filtered image. To avoid this problem, a low-pass filter $B(u,v)$ of cutoff frequencies $f_{c_1}$ and $f_{c_2}$ was used in [6] to supress the high frequency components in $G(u,v)$ as follows:

$$\hat{G}(u,v) = G(u,v)B(u,v) \tag{9}$$

where in this method $f_{c_1} = f_{c_2} = f_c$ was assumed. Since $\hat{G}(u,v)$ satisfies the symmetric property conditions of 2D-DFT, then, the spatial kernel $\hat{g}(n,m)$ can be obtained by employing the inverse 2D-DFT, $\mathcal{F}^{-1}$, on $\hat{G}(u,v)$ as:

$$\hat{g}(n,m) = \mathcal{F}^{-1}\left[\hat{G}(u,v)\right] \tag{10}$$

In order to fit the memory of a given hardware, the final normalized spatial enhancement kernel is obtained by cropping $\hat{g}(n,m)$ with a desired size of $\tilde{r} \times \tilde{r}$:

$$\tilde{g}\left(n - (\frac{r-\tilde{r}}{2}), m - (\frac{r-\tilde{r}}{2})\right) = \frac{\hat{g}(n,m)}{\sum\sum_{\frac{r-\tilde{r}}{2} \leq \hat{n},\hat{m} < \frac{r+\tilde{r}}{2}} \hat{g}(\hat{n},\hat{m})} \tag{11}$$

where $\frac{r-\tilde{r}}{2} \leq n,m < \frac{r+\tilde{r}}{2}$, and the size of $\tilde{g}$ is $\tilde{r} \times \tilde{r}$.

In the proposed scheme, unlike the work in [6], we use two Wiener deconvolution kernels to enhance the input image instead of using only one kernel as in [6]. Let $\tilde{g}_T$ and $\tilde{g}_\Omega$ denote Wiener deconvolution kernels corresponding to the text-like and background regions, respectively. We design the two kernels using two different cutoff frequencies $f_{c_T}$ and $f_{c_\Omega}$, where $f_{c_T} > f_{c_\Omega}$, in order to enhance the input image $I_u$ according to the text mask obtained in (7).

## 4 Experimental Results

The proposed text-like region enhancement method has been tested on a $120Hz$ Christie projector[1] with a piezo-electric actuator introducing a half-pixel shift in both the horizontal and vertical directions. A software-triggered RGB camera is positioned to capture the superimposed projection results. The proposed scheme has been tested on four images, namely, *Eyechart*, *Video Card*, *Combined Style* and *Mixed Content*, and the original images are shown in the first column of Figure 3. All the test images include different types of text-like and background regions. The *Eyechart* image has text of different scales and a background of gradually varying intensity. For the *Video Card* image, it has text regions of

[1] Christie Matrix StIM WQ simulation projector

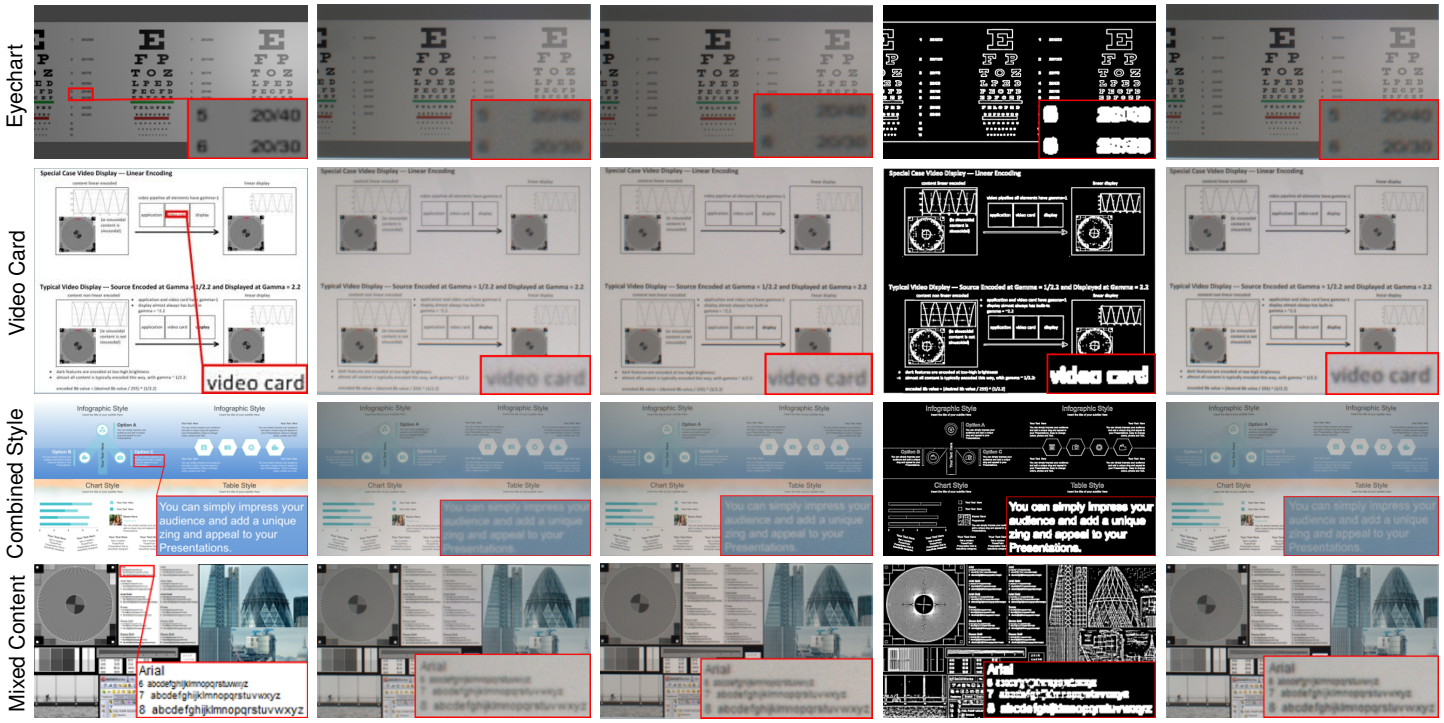


*Fig. 3:* Qualitative results of the proposed method (4th and 5th columns), projection without enhancement (2nd column) and Wiener Deconvolution [6] (3rd column) on four original images (1st column). Our proposed method was able to sharpen text more while avoiding over sharpening background for all test images.

different fonts, and it also has different graphs and charts that are sharp like text, such as sinusoidal waves, wheels and arrows. The *Combined Style* image contains text of different styles and rotation angles, as well as charts and tables, while the background and text have different colors. Finally, the *Mixed Content* image includes several text-like regions, such as text with multiple font sizes, lines and buildings.

The fourth column of Figure 3 shows the text masks of the proposed scheme on the test images in consideration. It is clear that the proposed scheme has been able to detect text-like regions of different fonts, styles and orientations. Figure 3 also provides a qualitative comparison among the proposed method, projection without enhancement (similar to [2]) and Wiener Deconvolution method in [6], where these methods have been tested on each image and the actual projection outputs were photographed by a camera from the projection screen. It is observed from the results that the proposed method offers a better visual resolution quality than that offered by the projection without enhancement and the method in [6], especially for text-like regions. For instance, the words "video card" in the second row of Figure 3 are more identifiable using the proposed scheme than the methods in comparison. In the selected region from the *Combined Style* image, the third row of Figure 3, the white text on a blue background are more clear using the proposed enhancement method compared with other methods. For the *Mixed Content*, the text-like regions have been generally enhanced by using the proposed method, for example, the word "Arial" of the selected captured region becomes more readable after applying the proposed scheme.

## 5 Conclusions

In this paper, we have presented a text-like region enhancement method for projector display using a new local dynamic range statistical thresholding. In this method, text-like regions have been obtained by computing the dynamic range of a spatial-neighborhood for every pixel in a given input image. Next, the local statistics of high dynamic range regions have been used to classify every pixel into text-like or background class. Two class-dependent Wiener deconvolution kernels of different cutoff frequencies have been used in order to sharpen the text-like regions higher than the background ones. Experimental results have been conducted on four challenging images and shown that the proposed scheme offers a better visual quality than that obtained by projection without enhancement and a recent state-of-the-art enhancement method.

## Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC-CRD), Christie Digital Systems Inc., and the Ontario Centres of Excellence (OCE-VIP2).

## References

[1] K. Hamada, M. Kanazawa, I. Kondoh, F. Okano, Y. Haino, M. Sato, and K. Doi, "A wide-screen projector of 4k x 8k pixels," in *Proc. SID Symp. Digest of Technical Papers*, vol. 33, no. 1, 2002, pp. 1254–1257.

[2] W. Allen and R. Ulichney, "Wobulation: Doubling the addressed resolution of projection displays," in *Proc. SID Symp. Digest of Technical Papers*, vol. 36, no. 1, 2005, pp. 1514–1517.

[3] N. Damera-Venkata and N. L. Chang, "Realizing super-resolution with superimposed projection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recogn.*, 2007, pp. 1–8.

[4] ——, "Display supersampling," *ACM Trans. Graph.*, vol. 28, no. 1, pp. 9:1–9:19, 2009.

[5] E. Barshan, M. Lamm, C. Scharfenberger, and P. Fieguth, "Resolution enhancement based on shifted superposition," *SID Symp. Digest of Technical Papers*, vol. 46, no. 1, pp. 514–517, 2015.

[6] A. Ma, A. Gawish, M. Lamm, A. Wong, and P. Fieguth, "Real-time spatial-based projector resolution enhancement," *Proc. SID Symp. Digest of Technical Papers*, vol. 49, no. 1, pp. 831–834, 2018.

[7] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761 – 767, 2004.

[8] M. Donoser and H. Bischof, "Efficient maximally stable extremal region (MSER) tracking," in *Proc. IEEE Conf. on Computer Vision and Pattern Recogn.*, vol. 1, 2006, pp. 553–560.

[9] K. H. X-C Yin, X-W Yin and H. Hao, "Robust text detection in natural scene images," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 970–983, 2014.

[10] M. Buta, L. Neumann, and J. Matas, "Fastext: Efficient unconstrained scene text detector," in *Proc. IEEE Int. Conf. on Computer Vision*, 2015, pp. 1206–1214.

[11] Y. Niblack, *An Introduction to Digital Image Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall Int. Inc., 1986.

[12] M. Smith and T. Kanade, "Video skimming for quick browsing based on audio and image characterization," Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-CS-95-186, July 1995.

[13] S. Antani, D. Crandall, and R. Kasturi, "Robust extraction of text in video," in *Proc. Int. Conf. on Pattern Recogn.*, 2000, pp. 831–834.

[14] M. R. Lyu, J. Song, and M. Cai, "A comprehensive method for multilingual video text detection, localization, and extraction," *IEEE Trans. on Circuits and Syst. for Video Technol.*, vol. 15, no. 2, pp. 243–255, 2005.

[15] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006.