Content-Adaptive Non-Stationary Projector Resolution Enhancement

by

Xiaodan Hu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
System Design Engineering

Waterloo, Ontario, Canada, 2019

Author's Declartion

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The following two papers are used in this thesis. I am the first author and have contributed to the design, implementation, experimentation, and writing of all the papers. The majority of contents of Chapter 4 has been taken from paper (1) and the majority of contents of Section 5.1 in Chapter 5 has been taken from paper (2).

1. X. Hu, M. A. Naiel, Z. Azimifar, I. Ben Daya, M. Lamm, and P. Fieguth. Text enhancement in projected imagery. *Journal of Computational Vision and Imaging Systems*, 4(1):3, Dec. 2018

Contributor	Statement of Contribution	
X. Hu (Candidate)	Conceptual design 70%	Writing and editing 81%
M. A. Naiel	Conceptual design 5%	Writing and editing 5%
Z. Azimifar	Conceptual design 5%	Writing and editing 5%
I. Ben Daya	Conceptual design 5%	Writing and editing 2%
M. Lamm	Conceptual design 5%	Writing and editing 2%
P. Fieguth.	Conceptual design 10%	Writing and editing 5%

2. X. Hu, A. Ma, A. Gawish, M. Lamm, and P. Fieguth. Motion detection in high resolution enhancement. *Journal of Computational Vision and Imaging Systems*, 3, 2017

Contributor	Statement of Contribution	
X. Hu (Candidate)	Conceptual design 80%	Writing and editing 84%
A. Ma	Conceptual design 2%	Writing and editing 10%
A. Gawish	Conceptual design 3%	Writing and editing 2%
M. Lamm	Conceptual design 5%	Writing and editing 2%
P. Fieguth.	Conceptual design 10%	Writing and editing 2%

Abstract

For any projection system, one goal will surely be to maximize the quality of projected imagery at a minimized hardware cost, which is considered a challenging engineering problem. Experience in applying different image filters and enhancements to projected video suggests quite clearly that the quality of a projected enhanced video is very much a function of the content of the video itself. That is, to first order, whether the video contains content which is moving as opposed to still plays an important role in the video quality, since the human visual system tolerates much more blur in moving imagery but at the same time is significantly sensitive to the flickering and aliasing caused by moving sharp textures. Furthermore, the spatial and statistical characteristics of text and non-text images are quite distinct. We would, therefore, assert that the text-like, moving and background pixels of a given video stream should be enhanced differently using class-dependent video enhancement filters to achieve maximum visual quality.

In this thesis, we present a novel text-dependent content enhancement scheme, a novel motion-dependent content enhancement scheme and a novel content-adaptive resolution enhancement scheme based on a text-like / non-text-like classification and a pixel-wise moving / non-moving classification, with the actual enhancement obtained via class–dependent Wiener deconvolution filtering. Given an input image, the text and motion detection methods are used to generate binary masks to indicate the location of the text and moving regions in the video stream. Then enhanced images are obtained by applying a plurality of class-dependent enhancement filters, with text-like regions sharpened more than the background and moving regions sharpened less than the background. Later, one or more resulting enhanced images are combined into a composite output image based on the corresponding mask of different features. Finally, a higher resolution projected video stream is conducted by controlling one or more projectors to project the plurality of output frame streams in a rapid overlapping way.

Experimental results on the test images and videos show that the proposed schemes all offer improved visual quality over projection without enhancement as well as compared to a recent state-of-the-art enhancement method. Particularly, the proposed content-adaptive resolution enhancement scheme increases the PSNR value by at least 18.2% and decreases MSE value by at least 25%.

Acknowledgements

I would like to thank my supervisors Prof. Paul Fieguth for his constant support during my masters studies. He set an incredible example for me as a researcher, teacher and mentor. Thank you both for all the guidance and support in my quest to become a researcher.

I would like to thank both Dr. Zhou Wang and Dr. John Zelek for serving as readers of my Masters thesis.

I would like to thank my co-authors Zohreh Azimifar and Mohamed Naiel for their consistent inputs and suggestions during my research, and their warm support as friends.

I would also like to thank the members of the Vision and Image Processing Lab. Witnessing their success motivated me. Thank you for inspiring me to do great research, to challenge myself and to make a difference. I enjoyed every moment working with you.

In addition, I would like to thank Mark Lamm and Christie Digital for providing me an excellent internship opportunity through which I have been able to experience industry oriented research.

Finally, I want to thank my parents and my boyfriend for their continuous support and encouragement throughout my years of study. They let me be me. Thank you.

Dedication

This is dedicated to the people I love.

Table of Contents

List of Tables x

List of Figures xi

1 Introduction 1

1.1 Motivation and Overview 1

1.2 Problem Statement and Objectives 3

1.3 Thesis Organization 3

2 Background 5

2.1 Shifted Superposition 5

2.2 Spatial-based Wiener Deconvolution Filtering 8

2.3 Text-like Region Detection 9

2.4 Conclusion 10

3 Problem Formulation 12

3.1 Text Enhancement Formulation 14

3.2 Motion Enhancement Formulation 16

3.3 Content-adaptive Enhancement Formulation 18

3.4 Conclusion 20

4 Projector-based Text Enhancement **22**
4.1 System Model . . . 23
4.2 Text Detection Methods . . . 23
4.2.1 Local Dynamic Range Statistical Thresholding . . . 25
4.2.2 Local Statistical Bimodality . . . 26
4.2.3 Bimodal Text Detection via Gray Pixel Counting . . . 28
4.3 Text Enhancement . . . 29

5 Projector-based Motion Enhancement **31**
5.1 Optical Flow-Based Motion Enhancement . . . 32
5.1.1 Motion Estimation using Optical Flow . . . 32
5.1.2 Kalman Filter . . . 34
5.1.3 Advanced Motion Estimation using Optical Flow and Kalman Filter 34
5.1.4 Scene Cut Detection . . . 35
5.1.5 Directional Blurring Filter . . . 37
5.2 Hypothesis Testing-Based Motion Enhancement . . . 37
5.2.1 System Model . . . 37
5.2.2 Motion Detection . . . 38
5.2.3 Motion Enhancement . . . 41

6 Content-adaptive Resolution Enhancement **42**
6.1 Introduction . . . 42
6.2 Non-Stationary Filtering . . . 44
6.3 Conclusion . . . 45

7 Experimental Results **47**
7.1 Datasets . . . 47
7.2 Text Enhancement Results . . . 49

7.3 Motion Enhancement Results 51
7.3.1 Optical Flow-Based Motion Enhancement Results 51
7.3.2 Hypothesis Testing-Based Motion Enhancement Results 56
7.4 Content-Adaptive Non-stationary Projector Resolution Enhancement Results 63
7.5 Summary 63

8 Conclusion **65**
8.1 Summary of Thesis and Contributions 65
8.2 Impact and Future Work 66

References **67**

List of Tables

7.1 Quantitative results of text enhancement method 49
7.2 Average mean square error for “spinning” video 54
7.3 Average mean square error for “moving lines” video 54
7.4 Quantitative results of hypothesis-based motion enhancement method . . . 58
7.5 Motion artifacts measurements 63

List of Figures

2.1 An overview of shifted superposition 6

3.1 Projector projection of sample image containing text-like regions 13
3.2 An overview block diagram of the proposed text enhancement scheme . . . 14
3.3 Moire artifacts produced by enhancement of high contrast imagery 16
3.4 Proposed moving content enhancement scheme 17
3.5 Block diagram of the proposed motion detection scheme 18
3.6 The proposed non-stationary content-adaptive enhancement scheme 19

4.1 An overview block diagram of the proposed text enhancement scheme . . . 24

5.1 Scene cut detection using ECR method 36
5.2 Proposed moving content enhancement scheme 38
5.3 Block diagram of the proposed motion detection scheme 39

6.1 The proposed non-stationary content-adaptive enhancement scheme 43

7.1 Sample Videos of the Visual Projection Assessment Dataset (VPAD) . . . 48
7.2 Overview of the Visual Projection Assessment Dataset (VPAD) 48
7.3 Qualitative results of the proposed text enhancement method 50
7.4 Flow visualization 52
7.5 Motion map calculated using proposed method and other methods 53
7.6 Qualitative results of optical-flow-based enhancement method 55

7.7 Scene cut detection applied on motion estimation 57
7.8 Qualitative results using directional blurring 58
7.9 Motion artifacts comparison 59
7.10 Qualitative results of hypothesis-based motion enhancement method 60
7.11 Qualitative results of non-stationary enhancement method 62
7.12 Qualitative results of content-adaptive non-stationary enhancement method 64

Chapter 1

Introduction

This thesis is aiming to improve the quality of projected imagery with a minimized hardware cost. This chapter reviews the research literature on projector resolution enhancement and discusses the weakness of currently available resolution enhancement methods. The study is carried out due to the concern that the features of text-like regions, moving regions, and background regions are significantly different. The edges of static text-like regions usually need to be distinct enough for better readability. However, moving regions are easy to be over-sharpened leading to highly distracting motion artifacts. That is, enhancing all kinds of regions equally will lead to some visual problems, and a content-adaptive enhancement method is desired in a real-world situation.

1.1 Motivation and Overview

High-resolution content projection systems are enjoying increasing popularity in consumer markets [3], with ultra-high resolution projectors having been available in theaters and amusement parks [4]. For example, the current projector display technology already has the capability to project 4k cinema content, producing brilliant, high-resolution visuals.

Although these high-resolution projectors have been available for some time, they are still costly. In 2017, the number of installed 4k projectors only comprised 17% of total screens worldwide. Also, at the same time, more and more films and videos are being captured using cameras capable of 4k, ultra-high-definition, or 8k resolution, and require high-resolution projection. However, the prevalent projectors cannot generally project such high-resolution videos. Therefore the use of one or more low-resolution projectors to display high-resolution content remains a highly attractive option.

The concept of using low-resolution projectors to project high-resolution images has been explored in the literature, and many approaches have been proposed [5, 6, 7, 8, 9]. We are motivated by the Wobulation method proposed by Allen and Ulichney [6]. Wobulation is a cost-effective method of increasing the resolution of digital projection systems. Wobulation method first generates multiple low-resolution sub-frames based on each frame of higher resolution image data, and then the sub-frames are overlaid subject to shifting by a fraction of a pixel. The sub-frames are displayed in rapid succession, thereby appearing as if they were projected simultaneously and superimposed, producing perceptible high-resolution content than images produced by unwobulated systems.

Later, a Shifted Superposition (SSPOS) method was proposed by Barshan [10] based on the Wobulation scheme [6]. In order to further improve the Wobulation method, SSPOS proposes an optimization method to generate two optimized low-resolution sub-images, based on the high-resolution target image.

Though Barshan [10] solved an optimization problem to find the sub-images, the use of a local Wiener filter is much simpler, if not entirely offering the same performance as image-dependent optimization. In particular, the Wiener filter can be used [11] as a deconvolution to compensate for optical aberration from the projector-lens systems.

However, although both SSPOS [10] and the Wiener deconvolution based enhancement [11] work well for still images, in videos both methods introduce temporal motion artifacts due to their inherent sharpening operation. These artifacts (e.g., flickering and aliasing) are associated with moving detailed texture, and time-nonstationarity / sequencing introduced by the superimposed projection. The artifacts may be present, but not obvious, in still images, however, they are highly distracting/intolerable in videos. A simple and effective method to reduce these artifacts is to apply blurring on the whole image but which is, of course, a frustrating solution in a method aiming to offer an increased resolution for moving and non-moving regions. In addition, effective filtering for text-like regions tends to create artifacts in other contents, and effective filtering for other image contents tends to under-enhance text-like regions. Thus, effective enhancement of text-like, motion and other imagery regions is of significant importance and represents significant added value for projector display systems.

What is more, though the content-adaptive enhancement scheme has been proposed [12], it does not provide the solution of how to classify a given video frame into different content classes, and how to combine different enhanced contents into a final enhanced frame.

1.2 Problem Statement and Objectives

Consequently, we have the following problems at hand:

1. The detection of text–like regions (Section 4.2)
2. The enhancement of text–like regions (Section 4.3)
3. The detection of moving regions (Section 5.2.2)
4. The enhancement of moving regions (Section 5.2.3)
5. The smoothing technique to avoid sharp transitions between different regions (Section 3.3)
6. A novel content-adaptive projector resolution enhancement scheme (Chapter 6)

The problem formulation will be developed in further detail in Chapter 3. Therefore, the objective of this thesis is to find a content-adaptive enhancement method that offers sharpening the text-like regions higher than the background ones, while avoids over-sharpening moving regions which may cause temporal motion artifacts. As such, text-like and moving regions in the final enhanced image can be all enhanced in an appropriate way to offer better visual quality for projected video frames than that of projection without enhancement.

1.3 Thesis Organization

The rest of this thesis is structured as follows.

Chapter 2 describes the relevant studies of projector enhancement, including an overview of text detection methods, the state-of-the-art projector resolution enhancement methods, as well as the hardware mechanism to accomplish higher resolution projection using low-resolution projectors.

Chapter 3 provides a brief description of the problem that this thesis targets.

Chapter 4 introduces a text enhancement scheme which consists of several novel methods for text detection, text enhancement filters.

Chapter 5 introduces two motion enhancement schemes, one resolves the motion artifacts problem by directionally blurring the moving regions, and another one enhances moving regions by applying a less sharpened filter.

Chapter 6 introduces a comprehensive enhancement scheme which both resolves the motion artifacts problem and makes text regions sharper. In this chapter, non-stationary filtering is used to combine different enhanced regions smoothly.

Chapter 7 introduces the dataset and shows the qualitative and quantitative experimental results of text enhancement, motion enhancement, and content-adaptive enhancement.

Chapter 8 concludes the work and describes the future works.

Chapter 2

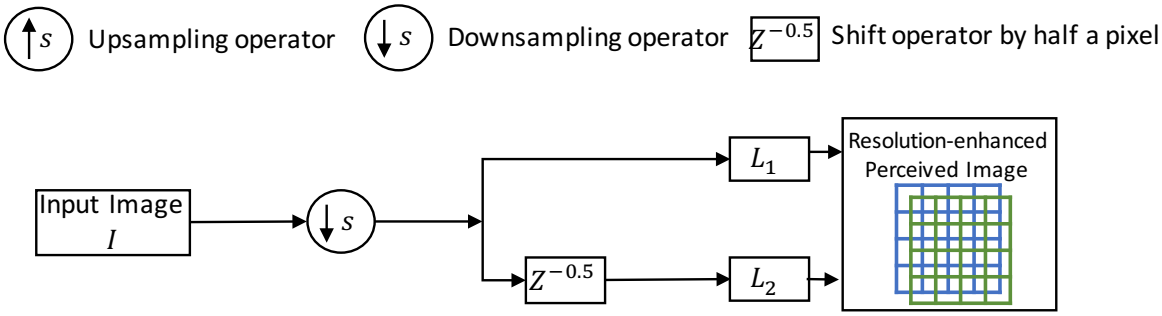
Background

This chapter introduces the relevant information required to understand the proposed content in this work. First, Section 2.1 reviews a picture of a projector system which produces higher resolution using a low-resolution projector. Next, Section 2.2 formalizes a local spatial-based Wiener deconvolution filter for resolution enhancement. Finally, a review of various text detection methods are in Section 2.3.

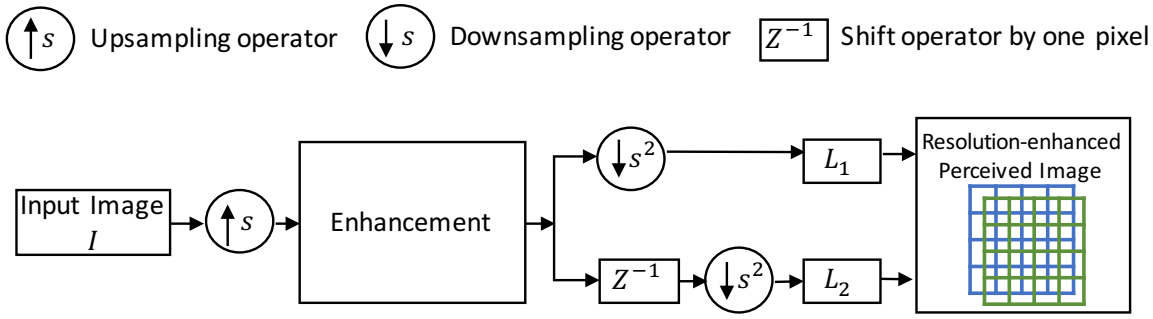
2.1 Shifted Superposition

The spatial resolution of a video can be increased by first generating two low-resolution frames from the high-resolution frame shifted by a fraction of a pixel, and then overlaying and displaying them within the retinal integration time to achieve higher perceived resolution [13]. Such a resolution enhancement approach has been an active research area [6, 14, 15, 16, 17, 18, 19, 20, 21]. For the case of single projector display, the shifted superimposed projection is achieved by wobulating the image [6], vibrating the entire display [18, 19] or overlapping multiple pixels [17]. Alternatively, such a resolution enhancement can also be achieved with a multi-projector setup by superimposing multiple projections from one or more machines [20, 21]. In this thesis, the proposed resolution enhancement scheme is based on the shifted superimposed projection shown in Figure 2.1 (b) to achieve a higher resolution projection since the single-projector projection setup is simpler than the multi-projector projection setup.

Figure 2.1 (b) is an overview of the improved shifted superposition scheme. The high-resolution image I is projected using two low-resolution images, L_1 and L_2, shifted by



(a) Shifted Superimposed Projection



(b) Improved Shifted Superimposed Projection

Figure 2.1: An overview of shifted superposition. (a) is a general shifted superposition scheme [6]. Given an input image I, two low-resolution sub-images, L_1 and L_2, are generated by first downsampling I and then shifting by half a pixel. Then L_1 and L_2 are superimposed in rapid succession to reconstruct a higher resolution image. (b) is an improved shifted superposition scheme [22]. In this scheme, the half-pixel shift component is replaced by an one-pixel shift component followed by a downsampling operators since implementing the one-pixel shift in a projector is much easier. In order to achieve the half-pixel shifting between L_1 and L_2, the image shifted by one pixel is downsampled by a downsampling operator with $s^2 = 2$. Also, an enhancement component is added to enhance the input image I.

half a pixel. The half-pixel-shifted L_1 and L_2 are obtained by first shifting the upsampled enhanced image $\hat{I}_u$ by one pixel and then downsampling by a factor of 2.

Let s and $1/s$ denote image upsampling and downsampling factors in both x and y directions, where $s > 1$. Further, let $\mathcal{P}(I, s)$ be a resampling function for an input image I by a resampling factor s in both x and y directions. Now, the two sub-images L_1 and L_2 are generated by first up-sampling I by a factor of s in both x and y directions to obtain I_u as follows:

$$I_u = \mathcal{P}(I, s) \tag{2.1}$$

Then, the upsampled image I_u can be enhanced by applying an enhancement function, let it be denoted as $\mathcal{Q}(I_u)$, to obtain $\hat{I}_u$ as

$$\hat{I}_u = \mathcal{Q}(I_u) \tag{2.2}$$

Next, the two enhanced low resolution images L_1 and L_2 can be obtained as follows:

$$L_1 = \mathcal{P}(\hat{I}_u, \frac{1}{s^2}) \tag{2.3}$$

$$L_2 = \mathcal{P}(\mathcal{Z}(\hat{I}_u, 1), \frac{1}{s^2}) \tag{2.4}$$

where $\mathcal{Z}(\hat{I}_u, 1)$ is a shifting operator for $\hat{I}_u$ by one pixels in both vertically and horizontally. From (2.3) and (2.4), the two low resolution sub-images are obtained by down-sampling $\hat{I}_u$ and its one-pixel shifted version by a factor of s^2 where $s^2 = 2$.

Then the two sub-images are superimposed on a given projection surface in rapid succession to reconstruct a higher resolution image that approximates the target image, where this process is called shifted superimposed projection [6]. In order to implement this process, the hardware configuration is designed to produce shifted superimposed projection that approximates projecting the high-resolution image. The superposition projection is implemented in hardware by optics and mechanics to diagonally shift the projected image to provide half-pixel shift. In this case, the original 60Hz ultra-high-definition video is reproduced by a 120Hz low-resolution video shown in two shifted pixel positions.

In the literature, several attempts have been carried out to design an enhancement function, $\mathcal{Q}$, that affects the superimposed projection quality, such as the methods in [10, 22]. However, these methods [10, 22] introduce motion artifacts due to over-sharpening moving regions and inadequate enhancement due to under-sharpening text regions. In this thesis, several enhancement functions $\mathcal{Q}$ are proposed in Chapter 4, Chapter 5 and Chapter 6.

2.2 Spatial-based Wiener Deconvolution Filtering

Wiener deconvolution filter is widely used for image deblurring and restoration [11, 23], and the possibility of applying the Wiener deconvolution filter to reconstruct images from projections has been pointed out by Klug [24]. The deconvolution filtering process is defined in the transform domain and attempts to minimize the impact of blurring due to the projector-lens system with a low signal-to-noise ratio (SNR) by reducing the least square error in prediction.

Point spread function (PSF) is a record of how much the image projected by a projector spreads/blurs an object of a single point [25]. Given the estimated projector's 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

$$G(u,v) = \frac{1}{H(u,v)} \left[\frac{|H(u,v)|^2}{|H(u,v)|^2 + \frac{1}{SNR}} \right] \tag{2.5}$$

where G and H are of size $r \times r$ and r is a constant factor that can be calculated empirically. Here, $1/H(f)$ is the inverse of the original projector-lens system, and $\mathrm{SNR}(f) = S(f)/N(f)$ is the SNR. When the noise is zero, the SNR becomes infinite, and the term inside the square brackets becomes 1. This means the Wiener deconvolution filter G becomes simply the inverse of the projector-lens system H. On the contrary, when the noise at certain frequencies increases, the SNR decreases, then the term inside the square brackets also drops from 1. That is, the Wiener deconvolution filter G attenuates frequencies dependent on their signal-to-noise ratio.

However, since the filtering process is defined in the transform domain, the filtering process results in high computational complexity. Recently, Ma *et al.* [22] introduced a local spatial kernel derived from the Wiener deconvolution filter to do the band-limited 2D convolution in the spatial domain.

It is shown in [22] that obtaining the spatial kernel $g(n, m)$ by directly applying the inverse 2D-DFT on (2.5) causes over-sharpening artifacts for filtered images. To avoid this problem, a low-pass filter $B(u, v)$ of cutoff frequency f_c was used in [22] to suppress the high frequency components in $G(u, v)$ as follows:

$$\hat{G}(u,v) = G(u,v)B(u,v) \tag{2.6}$$

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

$$\hat{g}(n,m) = \mathcal{F}^{-1}\big[\hat{G}(u,v)\big] \tag{2.7}$$

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 $\tilde{r} \times \tilde{r}$:

$$\bar{g}(\bar{n}, \bar{m}) = \hat{g}\left(\bar{n} + \frac{r - \tilde{r}}{2}, \bar{m} + \frac{r - \tilde{r}}{2}\right) \tag{2.8}$$

where $0 \leq \bar{n}, \bar{m} < \tilde{r}$, and the size of $\bar{g}$ is $\tilde{r} \times \tilde{r}$. The normalized version of $\bar{g}$ is obtained as follows:

$$\tilde{g}(\bar{n}, \bar{m}) = \frac{\bar{g}(\bar{n}, \bar{m})}{\sum \sum_{0 \leq \hat{n}, \hat{m} < \tilde{r}} \bar{g}(\hat{n}, \hat{m})} \tag{2.9}$$

The estimated deblurred image is obtained by filtering the blurred image with the local spatial-based Wiener deconvolution filter $\tilde{g}$.

2.3 Text-like Region Detection

During the exploration of resolution enhancement filters, on the one hand, the sharpening strength that is appropriate for text-like regions is too sharp for other regions (e.g., moving regions), resulting in additional artifacts. On the other hand, proper sharpening strength for other regions tends to under-enhance the text-like regions. Therefore, the text-like regions and non-text-like regions should be enhanced differently, and an efficient and effective text-like region detector that can classify each pixel into text or non-text classes is required.

An efficient and effective text-like detection is a challenging visual recognition problem [26, 27]. Furthermore, a good text-like detection method for projector projection is even more challenging since it should also consider the bandwidth and memory requirements of the projector hardware. Current available image text detection methods can be classified into three categories: optical-character-recognition-based text detection methods, connected component-based text detection methods, and image-thresholding-based text detection methods.

First, among the traditional text detection methods, Optical character recognition (OCR) is widely used by many research works [28, 29, 30, 31, 32, 33, 34]. OCR has the capability to do text detection and text extraction, and has been steadily evolving during its history. OCR produces a ranked list of candidate characters by comparing to glyph prototypes. Some OCR methods compare the image with the stored glyph pixel by pixel while some first decompose the image to features and then compare the extracted

features with the stored glyph features to find the closest match. However, OCR assumes that the characters are typically monotone on fixed backgrounds, which cannot adapt to more complicated scenarios such as variations in font, background, color, and light [27]. Moreover, in this thesis, we only need a classifier to classify each pixel into a text-like or a non-text-like class, while OCR-based methods provide not only text detection but also text recognition, and firmly relies on the pre-processing, post-processing and optimization steps such as character segmentation and de-skew to make text horizontal or vertical, grammar checking [35]. Since character recognition is unnecessary in our task (only the text-like regions classification is required), we are looking for a more efficient text detection method.

Connected component-based text detection methods have gained high attention in many studies [36, 37, 38, 39]. For instance, Matas *et al.* [36] introduced the Maximally Stable Extremal Regions (MSER) method which, as a feature extractor, partitions an input image into a number of co-variant indivisible components, and followed by classifying each component as text or background. The extremal regions are obtained by applying a series of thresholds and thresholding the image at each level. The extremal regions are closed under a continuous transformation of image coordinates or a monotonic transformation of image intensities. Thus, the extremal regions are stable and invariant to affine transformations of image intensities and co-variant to adjacency preserving transformations on the image domain. However, it is known that MSER is sensitive to noise and image blurring since it depends on successful edge detection, and it needs to save the candidate regions under different thresholding levels, which increases the burden of the projector hardware [40].

The image-thresholding-based text detection methods can create a binary map classifying each pixel into text-like or non-text-like class. Many image-thresholding-based text detection methods have been introduced [41, 42, 43, 44]. For instance, in [42], a local horizontal differential filter and a thresholding scheme were performed in order to find vertical edges in an image. This method, however, lacks representing the 2D structures within the text regions. Later in [43], 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.

2.4 Conclusion

This chapter introduces the background knowledge that will be used in the proposed content-adaptive resolution enhancement scheme. First, Section 2.1 introduces shifted

superposition, a method to project higher resolution image using a low-resolution projector. The shifted superposition superimposes two sub-images on a given projection surface with half-a-pixel shift in rapid succession to reconstruct a higher resolution image that approximates the target image. Then, Section 2.2 describes a local spatial-based Wiener deconvolution filter for resolution enhancement, where the sharpening strength can be adjusted by changing the cutoff frequency of its low-pass filter. Finally, various text detection methods are reviewed in Section 2.3, suggesting the necessity of finding a new text detection method efficient and effective for projector display.

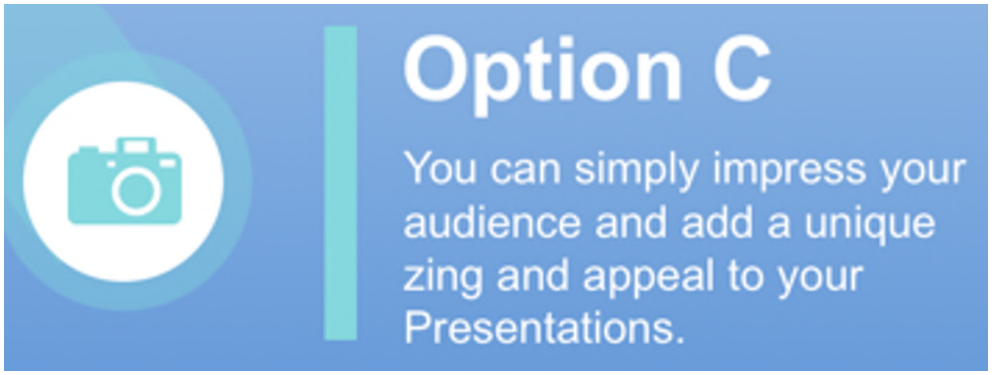
Chapter 3

Problem Formulation

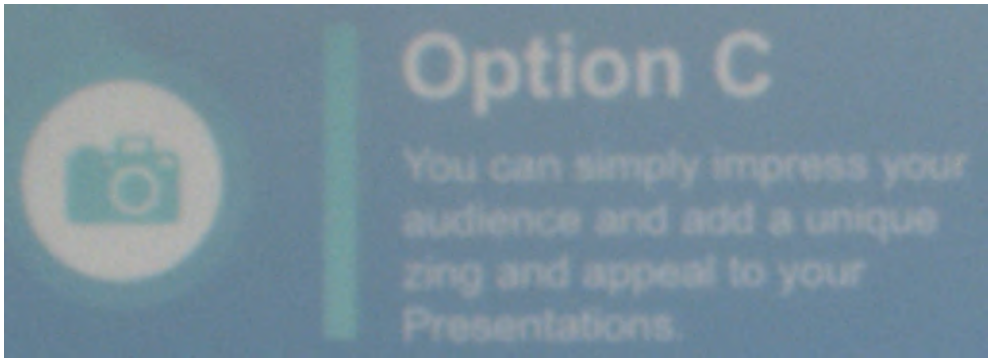
Chapter 1 has discussed the problem that current projector resolution enhancement methods do not consider different sharpening requirements of contents with distinct features. For instance, using the same sharpening strength for text-like regions are likely to create motion artifacts in moving contents while using a less sharpening strength suitable for moving contents will under-enhance the text-like regions. Hence, a content-adaptive enhancement method that can sharpen different features of a given image in different levels is desired.

The shifted superimposed projection [6] introduced in Section 2.1 can be used to project higher-resolution images using low-resolution sub-images. The Wiener deconvolution enhancement method in [22] introduced in Section 2.2 enhances the projector resolution by a local spatial Wiener deconvolution filter. The Wiener deconvolution filter can be used to enhance the image with different strength. Hence the content-adaptive enhancement method can adopt the shifted superimposed projection method to provide higher resolution projection and a set of Wiener deconvolution filters to enhance different regions. The focus of this thesis will be on content detection and content-dependent enhancement filters.

The rest of the chapter is structured as follows. Section 3.1 formulates the text enhancement problem where text readability should be improved. Section 3.2 formulates the motion enhancement problem to avoid motion artifacts while still enhancing moving regions. Section 3.3 formulates the comprehensive content-adaptive resolution enhancement which will consider both text readability and motion artifacts. Section 3.3 provides a method to combine different enhanced contents together to achieve content-adaptive resolution enhancement.



(a) Original Image



(b) Projected Image Without Enhancement

Figure 3.1: An sample image containing text-like regions (a) is compared with the projector projection of the sample image (b), which shows the necessity of text enhancement. Note that (b) is captured from the screen using a camera. The blurry text in the text-like regions in the projected image without enhancement is illegible while the background regions are more acceptable since background regions do not contain any important information.

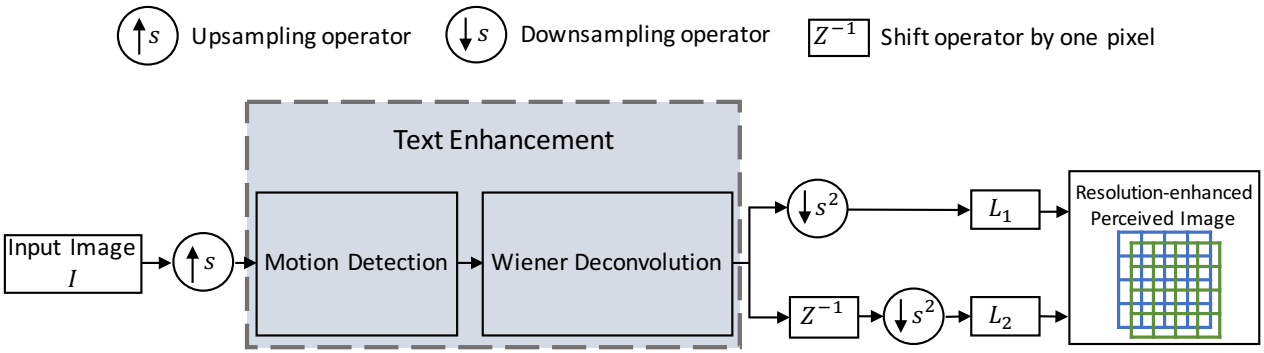


Figure 3.2: An overview block diagram of the proposed text enhancement scheme. The upsampling, downsampling and shifting components are used to produce the two sub-frames for pixel-shifted superimposed projection. The text enhancement includes two components, text-like detection and Wiener deconvolution. A text mask M_T is calculated in the text-like detection component to indicate the location of text-like regions. The enhanced image $\hat{I}_u$ is obtained via the Wiener deconvolution component where text-like Wiener deconvolution kernel g_T and background kernel g_Ω are applied to text-like and background regions, respectively. After the enhanced image is obtained, shifted superimposition method will be used for projection, where s is an upsampling factor in both x and y directions, and L_1 and L_2 are two downsampled sub-images generated with and without one pixel shift, respectively.

3.1 Text Enhancement Formulation

Since the 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 shown in Figure 3.1, given a sample image I (a), the text regions in the projected image (b) without enhancement are blurred by the projector-lens system. The small white paragraph in (b) is much darker and blurrier than that in the original image. For example, the two letters “e” and “a” both have similar structure, and the strokes are both dense, making them undistinguished in projected image (b). Furthermore, the boundaries of the camera icon and green vertical bar on the right are ambiguous. The camera icon, after projection, even looks like a handbag, which is misleading.

Nevertheless, as introduced in Chapter 1, although recent projector resolution enhancement methods are able to increase the perceived image resolution and also perform res-

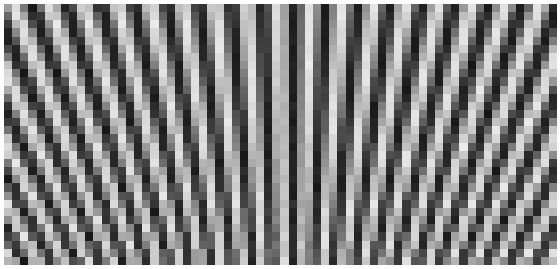
olution enhancement [6, 10, 22], they do not consider different sharpening requirements of text-like regions and non-text-like regions. For instance, in exploring possible Wiener deconvolution filters [22], it was clear that effective filters for text tended to create artifacts in other content, and effective filters for imagery tended to under-enhance text. Hence, considering the importance of the sharpness of text regions and the failure of state-of-the-art enhancement methods, a content-adaptive enhancement method that can sharpen different features of a given image in different levels is desired. As a result, we have two basic problems at hand:

1. The detection of text–like regions (Section 4.2)
2. The enhancement of text–like regions (Section 4.3)

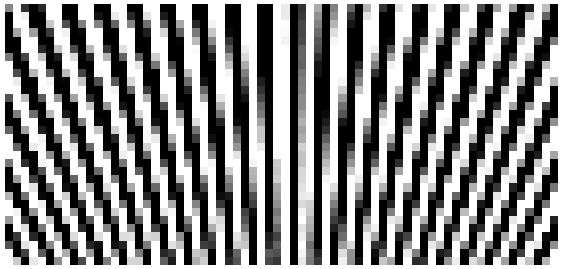
Figure 3.2 is an overview of the proposed text enhancement scheme. In order to do the text enhancement, the location of text-like regions should first be known and an enhancement filter suitable for text-like regions should be obtained. Therefore, the objective of text enhancement is to find an effective text-like region detection method to obtain a text-like mask indicating the locations of text-like regions and effective enhancement filters to enhance the image, respectively, where the text regions can be enhanced more than the background.

The detection of text–like regions: The text-like region detection methods are based on two natures of text-like regions. The first one is that they are always of higher contrast than the other regions. Thus, the local dynamic range of the pixel values can be used as a measurement of image contrast. The second one is that the distribution of pixel values in text-like regions are always bi-modal. Therefore, a clustering method can help to determine the bimodality of a given region. In Section 4.2, three text detection methods are proposed based on these two characteristics of text regions, respectively.

The enhancement of text–like regions: After text-like regions being detected, a sharper enhancement filter should be applied to the text-like regions and a less sharp filter on other regions. As introduced in Section 2.2, Ma *et al.* [22] proposed a Wiener deconvolution filtering-based resolution enhancement method where the sharpening strength of the Wiener deconvolution filter can be adjusted by changing the cutoff frequencies.



Original Image



Sharpened Image

Figure 3.3: Enhancement of high contrast imagery (left) can produce Moire artifacts (right), which can become dizzying / spinning distractions when moving. The Moire pattern in the right image is one of the motion artifacts caused by sharpening the original image.

3.2 Motion Enhancement Formulation

Even though most current state-of-the-art resolution enhancement methods work well for still images, in video these methods introduce temporal motion artifacts due to their inherent sharpening operation [19].

In Figure 3.3, the right image shows the sorts of Moire artifacts associated with over–sharpening by using the Wiener deconvolution-based method in [22]. Tests on super-imposed projection display have made very clear that effective filters for static contents, particularly high-contrast contents (such as text), tended to create badly-aliased artifacts in moving regions; similarly filters effective for motion tended to under-enhance / blur other high-contrast contents. Since the human visual system is relatively insensitive to the blur of moving objects [45], and the super-imposed projection creates the challenge of potential aliasing primarily for moving content, clearly the appropriate response is some sort of motion-dependent enhancement, which is the focus of this thesis. As a result, we have two basic problems at hand:

1. The detection of moving regions (Section 5.2.2)
2. The enhancement of moving regions (Section 5.2.3)

In order to achieve this goal, in this thesis, first, the motion regions should be detected to indicate the location of moving regions that need to be enhanced with less strength. And then, an enhancement filter for moving regions and an enhancement filter for non-moving

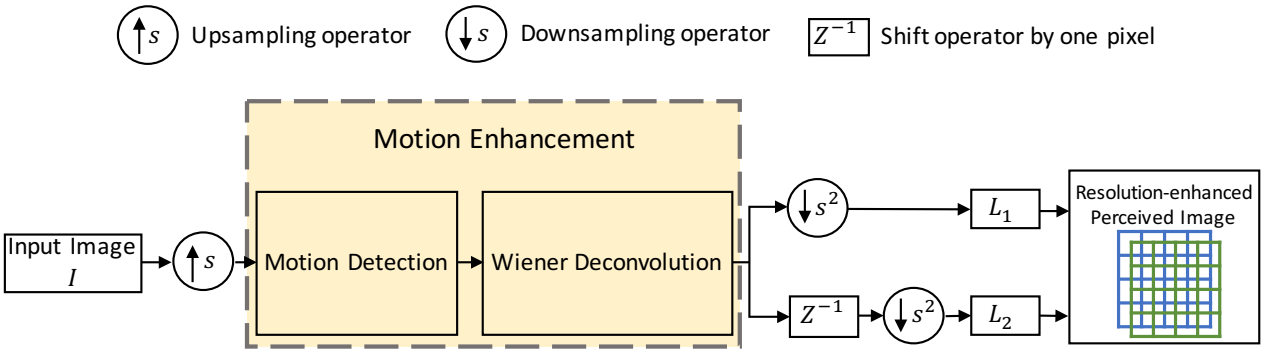


Figure 3.4: Proposed moving content enhancement scheme, which is a direct counterpart of Figure 3.2. Rather than sharpening text-like regions with a strength more than that of non-text-like regions, the motion enhancement scheme will sharpen moving regions less than that of non-moving regions.

regions should be obtained so that the moving and non-moving regions can be enhanced differently. Therefore, similar the the text enhancement scheme shown in Figure 3.2, a novel motion-dependent visual enhancement scheme in projector-based systems is proposed shown in Figure 3.4. The input video frame will go into the motion enhancement part to be enhanced based on moving and non-moving regions. Then, the enhanced image will be split into two low-resolution images by using shifted superimposed projection introduced in Section 2.1 to achieve higher resolution enhancement.

The detection of moving regions: In the motion enhancement part of this scheme, a robust motion detection, shown in Figure 3.5, is proposed to segment the moving regions from the background ones. This figure in this section is to give an overview of the proposed motion detection method, and the details inside the figure will be explained later in Section 5.2.2. The motion detection method takes β consecutive video frames as input and output a binary motion mask where white means moving regions. In order to distinguish moving pixels from non-moving ones, a motion threshold is obtained based on the statistics of frame differences. The idea behind it is that, given a pixel, if the pixel value varies significantly among the pixels in the same location in other frames, then say this pixel is moving. The threshold calculated in this figure is to determine whether the pixel values vary large enough to classified as moving. Otherwise, it will be classified as non-moving.

The enhancement of moving regions: Since moving regions tends to create artifacts when the enhancement strength is strong, the moving regions prefer an enhancement with

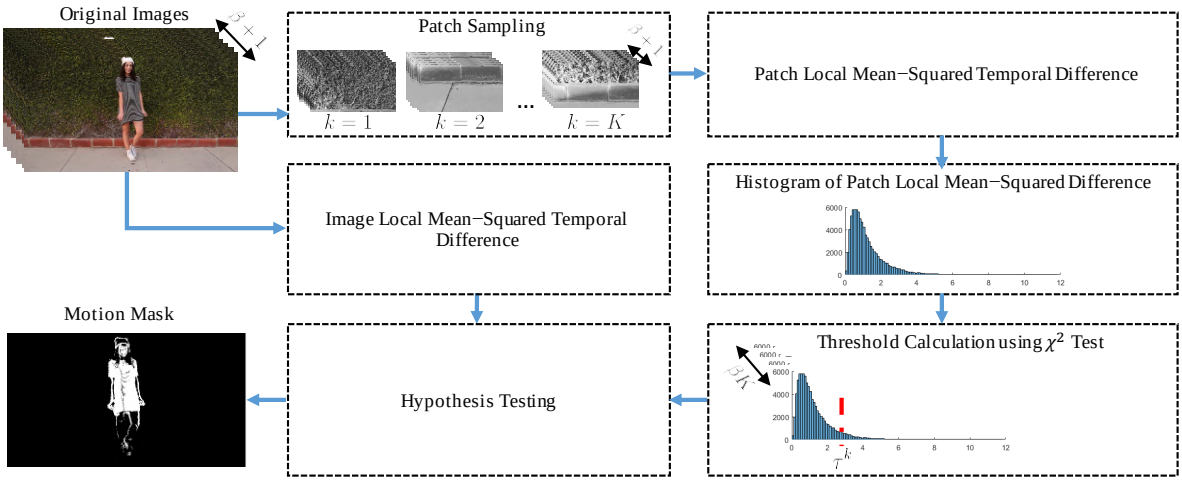


Figure 3.5: Block diagram illustration of the proposed motion detection scheme. This hypothesis-testing-based motion detection method takes multiple consecutive frames and considers the regions having relatively significant change among the input frames (the change of the brightness of pixels in these regions is greater than that of the majority of all the pixels) to be moving regions. The output is the motion mask where white means moving pixels and black otherwise. The concepts in this figure will be developed in further detail in Section 5.2.2.

a less sharpening strength. After motion regions are detected, a less sharpened and a more sharpened enhancement filter are applied to moving and non-moving regions, respectively, resulting in better visual quality. The first solution can be enhancing the regions based on the moving velocities. When the movement of a region is large enough, the region will be sharpened with less strength. In Section 5.1, an optical flow-based parameter selection technique is proposed in order to find the best sharpening levels to enhance the image while avoiding severe motion artifacts. A second solution can be enhancing the moving regions based on a motion mask indicating whether there is a motion or not. In Section 5.2, an hypothesis testing-based motion enhancement is proposed.

3.3 Content-adaptive Enhancement Formulation

The Wiener deconvolution filtering has been widely used to enhance the spatial resolution of images [46, 22]. However, these methods only have one enhancement kernel. The enhancement filter which can make the text sharper, however, will make moving detailed

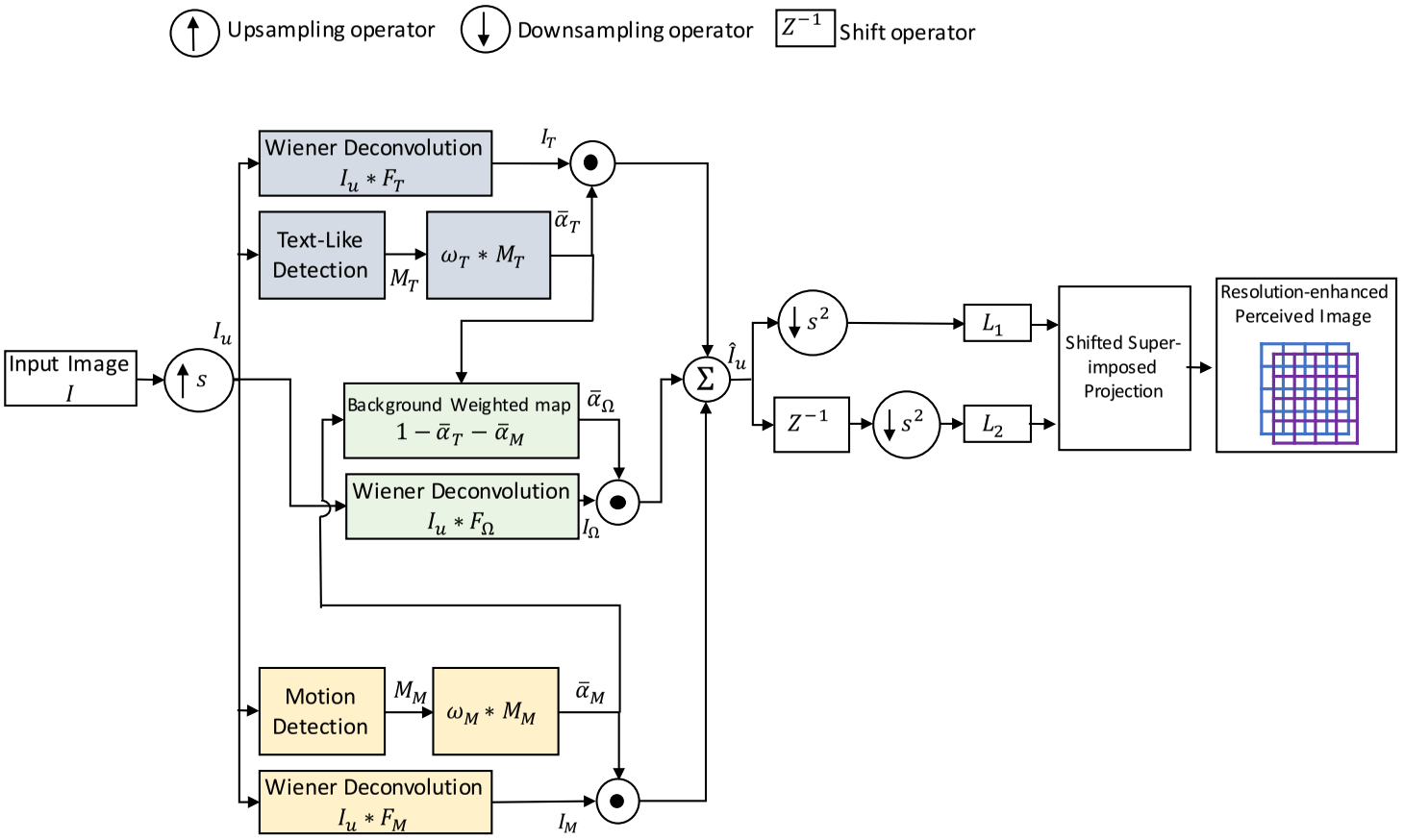


Figure 3.6: The proposed non-stationary content-adaptive enhancement scheme, which is building on the text and motion enhancement schemes shown in Figure 3.2 and 3.4. The colored components represent the main focuses of this thesis. The scheme includes a text enhancement component (grey), a motion enhancement component (yellow) and a background enhancement component (green) using different enhancement kernels. Each component can lead to the improvement of the perceptual quality of images based on the feature of the contents. The content detection methods, the selection of the corresponding enhancement kernel and the fusion of enhanced images are the focus in this work. The concepts will be developed in detail in Chapter 6.

textures worse. In the previous two sections, 3.1 and 3.2, the problems of text under-sharpening and motion over-sharpening have been formulated. Enhancing text and motion respectively is the possible solution. That is, when projecting videos contain text-like regions, a sharper enhancement filter will be used to sharpen the text regions and a less sharp filter will be used to sharpen the background. Similar to moving regions, a less sharp filter will be applied to moving contents and a stronger one for the background.

However, in real videos, much of the video content consists of both text-like regions and moving regions. Hence, a more comprehensive projector resolution enhancement is to enhance the text-like regions, moving regions and background simultaneously with different sharpening strengths. Nevertheless, tests on projection indicate that there will be an obvious boundary between more-sharply-enhanced text-like regions and less-sharply-enhanced moving regions. The combination of enhanced text-like regions and moving regions is expected to be more smoothly. As a result, we have a basic problem at hand:

1. A smooth combination of differently enhanced text-like and moving regions. (Section 6)

Figure 3.6 is the proposed projector-based content-adaptive resolution enhancement scheme. This figure provides a visual illustration of the concept of content-adaptive resolution enhancement which contains a series of components, and each component in this figure will explained in further detail in Section 6. The system includes text enhancement component (shown as grey boxes), motion enhancement component (shown as yellow boxes) and background enhancement component (shown as green boxes) using different Wiener deconvolution enhancement filters. Each component can contribute to the improvement of the perceptual quality of images. Then all the outputs from the text, motion and background enhancement components will be fused to generate the final enhanced composite image and fed to the projector for projection. As such, text-like and moving regions in the final enhanced image are all enhanced in an appropriate way.

3.4 Conclusion

In this Chapter, the problems of projector resolution enhancement have been formulated. Section 3.1 formulates the text enhancement problem. Since effective filters for text tended to create artifacts in other content, and effective filters for imagery tended to under-enhance text, text regions should be enhanced more than other regions. Therefore, effective and efficient text-like regions detector and text enhancement filter are required. Section 3.2

formulates the motion enhancement problem. Since effective filters for sparse static text-like contents tended to create aliasing artifacts in fine-detailed moving regions, the moving regions are expected to be enhanced with less strength than other regions. Hence, a moving regions detector and a moving regions enhancement filter is desired. Section 3.3 formulates the problem for comprehensive content-adaptive resolution enhancement, where text readability needs to be increased and motion artifacts should be avoided. It can be accomplished by sharpening different features of a given image at different levels. Particularly, text-like, moving and other regions should be first detected, then differently enhanced and combined to achieve content-adaptive resolution enhancement.

Chapter 4

Projector-based Text Enhancement

Since the 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.

However, as introduced in Chapter 1, although recent projector resolution enhancement methods are able to increase the perceived image resolution and also perform resolution enhancement [6, 10, 22], they do not consider different sharpening requirements of text-like regions and non-text-like regions. For instance, in exploring possible Wiener deconvolution filters [22], 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. Hence, a content-adaptive enhancement method that can sharpen different features of a given image in different levels is desired. As a result, we have two basic problems at hand:

1. The detection of text–like regions (Section 4.2)
2. The enhancement of text–like regions (Section 4.3)

In this chapter, 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 [1]. 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 better visual quality for the projected content. It is shown from the sample results in Figure 3.1 that unlike the method in [22], 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 better quality than that offered by projection without enhancement as well as the recent state-of-the-art enhancement method [22].

4.1 System Model

As shown in Figure 4.1, the proposed scheme consists of three main parts: text detection, 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 4.2, 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 previously in Section 2.2. The combined enhanced up-sampled image $\hat{I}_u$ of size $\hat{N}_1 \times \hat{N}_2$ is then obtained by adding the weighted images I_T and I_Ω using the corresponding weighted masks.

As in [22], 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 [10], 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 $\tilde{N}_1 = \lfloor \hat{N}_1/s^2 \rfloor, \tilde{N}_2 = \lfloor \hat{N}_2/s^2 \rfloor$. Finally, similar to [10] the two sub-images are superimposed to project perceived high-resolution contents.

4.2 Text Detection Methods

The first step to approach text enhancement is to develop a robust and efficient text detection technique to localize a wide variety of text-like regions of different shapes, colors, fonts, styles, sizes, and orientations. In this section, three proposed text-like region detection methods are introduced.

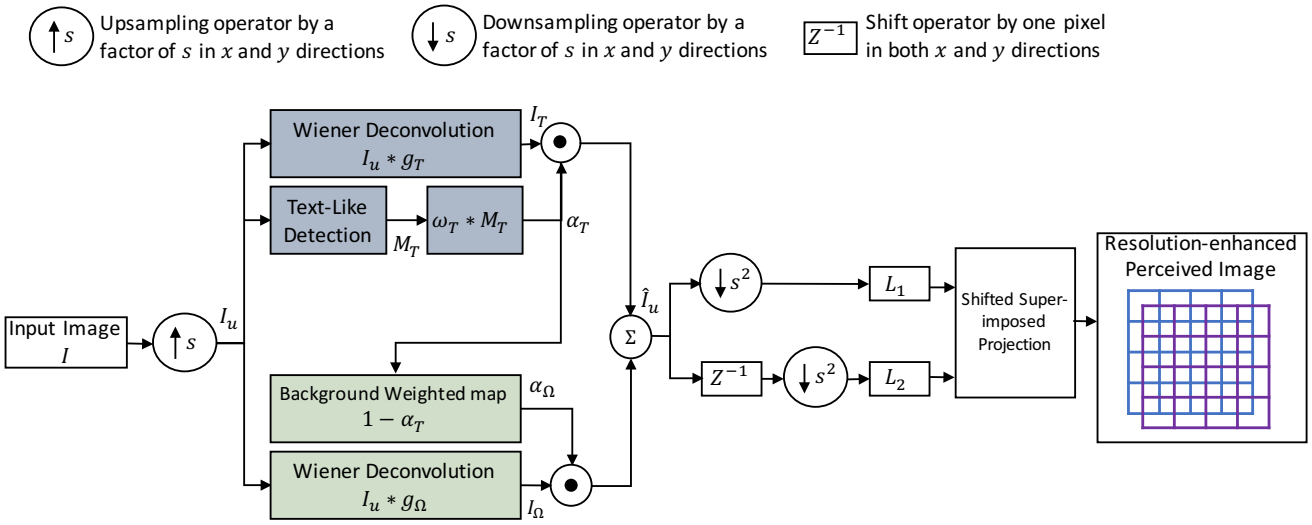


Figure 4.1: An overview block diagram of the proposed text enhancement scheme. The upsampling, downsampling and shifting components are used to produce the two sub-frames for pixel-shifted superimposed projection. The text enhancement includes two components, text-like detection and Wiener deconvolution. The text mask M_T is calculated in the text-like detection component. The resulted enhanced image $\hat{I}_u$ is obtained in Wiener Deconvolution component where text-like Wiener deconvolution kernel g_T and background kernel g_Ω applied to text-like and background regions, respectively.

4.2.1 Local Dynamic Range Statistical Thresholding

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 the local dynamic range to classify pixels into either text or non-text classes.

Since the color of text-like regions and their background should be distinct enough to ensure the text-like regions to be recognized, we assume that text-regions are of high-contrast. Based on this assumption, a region that is of high-contrast is considered as a text-like region. We propose to use the difference $D(i)$ between local maximum and minimum pixel values of the i^{th} local sliding window $\aleph_i$ of size $k \times k$ to represent the level of contrast of a given input image I, where i is the center of $\aleph_i$, $i = \Big[(1,1),(1,2),...,(\hat{N}_1,\hat{N}_2)\Big]$, $\hat{N}_1$ and $\hat{N}_2$ are the number of rows and columns of input image I. The local dynamic range $D(i)$ of the i^{th} pixel in I can now be obtained as

$$D(i) = \max_{j \in \aleph_i} (I(i,j)) - \min_{j \in \aleph_i} (I(i,j)) \tag{4.1}$$

Thus, text-like regions which is of high-contrast is having a higher dynamic range $D(i)$. The reason to use a sliding window is that a local window with an appropriate size can extract local features such as edges. Moreover, the contrast of a region is relatively a local comparison of pixel values within the region. Then, in order to determine whether a region is of high-contrast, a threshold needs to be applied to the local dynamic range $D(i)$ so that when $D(i)$ is greater than the threshold, the i^{th} pixel is considered to belong to text-like class.

However, $D(i)$ contains some background noise due to image compression. 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 τ_1 as follows:

$$\tau_1 = \{d^* \in D : h_{d^*} < h_d, d^* \neq d, d \in D\} \tag{4.2}$$

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

$$\tilde{D}(i) = \begin{cases} \hat{D}(i), & \hat{D}(i) > \tau_1 \\ \tau_1, & \text{Otherwise} \end{cases} \tag{4.3}$$

Then, a final threshold for $\tilde{D}$ can be computed. We proposed that in a given sliding window $\aleph_i$, when the dynamic range $\tilde{D}$ at the center is higher than most of the dynamic range values within the sliding window (average plus deviation), pixel i is considered as high-contrast, and thus text-like. The local statistics of the i^{th} element of the thresholded dynamic range $\tilde{D}$ are obtained as

$$\mu(i) = \frac{1}{k^2} \sum_{j \in \aleph_i} \left(\tilde{D}(j) \right) \tag{4.4}$$

$$\sigma(i) = \sqrt{\frac{1}{k^2} \sum_{j \in \aleph_i} \left(\tilde{D}(j) - \mu(i) \right)^2} \tag{4.5}$$

where μ and σ are the mean and the standard deviation of $\tilde{D}$. The final threshold τ_2 for $\tilde{D}$ is obtained as

$$\tau_2(i) = \mu(i) + \sigma(i) \tag{4.6}$$

The mask of text-like regions, M_T, is obtained by applying threshold τ_2 on $\tilde{D}$ as

$$M_T(i) = \begin{cases} 1, & \tilde{D}(i) > \tau_2 \\ 0, & \text{Otherwise} \end{cases} \tag{4.7}$$

where 1 represents text-like regions and 0 otherwise. This method performs well on our test images, however, the histogram operation in Equation (4.4) is not hardware friendly, and thus, we are looking for a simpler but also effective text-like region detection method.

4.2.2 Local Statistical Bimodality

Based on our observation of the feature difference between text-like regions and background, the pixel grey values in the text-like regions or background are similar and while the pixel grey values between text-like regions and background are quite different. Hence, we assume that the text-like regions behave in a bimodal manner. In order to test the bimodality of a region centered at the i^{th} pixel in a given image or color channel, I of size $N_1 \times N_2$, we first obtain the average, μ, and the standard deviation, σ, of the $\hat{N}_1 \times \hat{N}_2$ neighborhood pixels around the i^{th} location as follows:

$$\mu(i) = \sum_{j \in \aleph_i} \frac{I(j)}{|\aleph_i|} \tag{4.8}$$

$$\sigma(i) = \sqrt{\sum_{j \in \aleph_i} \frac{1}{|\aleph_i| - 1} \Big(I(j) - \mu(i) \Big)^2} \tag{4.9}$$

where $i = [(1,1),(1,2),...,(N_1,N_2)]$, N_1 and N_2 are the number of rows and columns of image I, $\hat{N}_1$ and $\hat{N}_2$ are the number of rows and columns of the local neighborhood $\aleph_i$, and $|\aleph_i|$ is the cardinality of the set $\aleph_i$. Then, the value of the average $\mu(i)$ corresponding to the i^{th} location is compared to the values of the neighborhood pixels at this location to categorize them into low group, $\aleph_i^-$ when a pixel value is less than $\mu(i)$ or high group, $\aleph_i^+$, otherwise:

$$\aleph_i^+ = \big\{ j | j \in \aleph_i, I(j) \geq \mu(i) \big\} \tag{4.10}$$

$$\aleph_i^- = \big\{ j | j \in \aleph_i, I(j) \leq \mu(i) \big\} \tag{4.11}$$

Next, the pixel values of the high and low groups are used to compute the average and standard deviation of each group as follows:

$$\mu_{high}(i) = \sum_{j \in \aleph_i^+} \frac{I(j)}{|\aleph_i^+|} \tag{4.12}$$

$$\sigma_{high}(i) = \sqrt{\sum_{j \in \aleph_i^+} \frac{1}{|\aleph_i^+| - 1} \Big(I(j) - \mu_{high}(i) \Big)^2} \tag{4.13}$$

$$\mu_{low}(i) = \sum_{j \in \aleph_i^-} \frac{I(j)}{|\aleph_i^-|} \tag{4.14}$$

$$\sigma_{low}(i) = \sqrt{\sum_{j \in \aleph_i^-} \frac{1}{|\aleph_i^-| - 1} \Big(I(j) - \mu_{low}(i) \Big)^2} \tag{4.15}$$

For a pixel to be classified as high-contrast, its variance $\sigma(i)$ given by the expression in (4.9) should be much larger than the average of the corresponding $\sigma_{high}(i)$ and $\sigma_{low}(i)$ and given by the expressions in (4.13) and (4.15), respectively. Otherwise, the pixel would be considered as belonging to the non high-contast class. We call this test as a bimodal test. To perform this bimodal test, a local threshold can be obtained for every pixel as

$$\Gamma(i) = (\sigma(i) - \lambda) - \left(\frac{\sigma_{low} + \sigma_{high}}{2} \right) \tag{4.16}$$

where λ is a constant obtained empirically for a given type of input color channel or other type of features such as image mask. The mask of the high-contrast regions, M_T, is obtained based on the threshold Γ as

$$M_T(i) = \begin{cases} 1, & \Gamma(i) > 0 \\ 0, & \text{Otherwise} \end{cases} \tag{4.17}$$

This text-like region detection method is based on the local statistical bimodality of a given region, which is quicker than the method proposed in previous section with similar results. However, this method is based upon the λ which is a constant obtained empirically. λ varies when the test images change greatly (e.g., different noise level and image quality). Therefore, a more robust thresholding method based on the same assumption of bimodality is desired.

4.2.3 Bimodal Text Detection via Gray Pixel Counting

As discussed in Section 4.2.2, the text-like regions are assumed to follow bimodal distribution. That is, the pixel grey values belong to text-like class should be close while pixel grey values of background class should also be close, but the grey value difference between text-like pixels and background pixels should be large. For a pixel i to be classified as text-like, the distribution of its neighbour pixel values should be bimodal. That is, after categorizing its neighbour pixel values into the high group, $\aleph_i^+$, and the low group, $\aleph_i^-$ (see Equation (4.10) and (4.11)), most of its neighbour pixel values are distributed around the average of high group, $\mu_{high}(i)$, and the average of low group, $\mu_{low}(i)$ (see Equation (4.13) and (4.15)), and few of their pixel values are in between. The reason is that if the pixel belongs to text-like regions, then most of its neighbour pixels are either foreground or background, and due to the characteristic that text-like regions usually have sharp edges, there are few pixels located in the transition regions between foreground and background. Otherwise, the pixel would be considered as belonging to the non-text-like class.

Define transition group S_i as the collection of the neighborhood pixels at location i where their pixel values are between $\mu_{high}(i)$ and $\mu_{low}(i)$:

$$S_i^t = \left\{ j | j \in \aleph_i, \mu_{low}(i) \leq I(j) \leq \mu_{high}(i) \right\} \tag{4.18}$$

Define non-transition group S_i as the collection of the neighborhood pixels at location i where their pixel values are larger than $\mu_{high}(i)$ or less than $\mu_{low}(i)$:

$$S_i^n = \left\{ j | j \in \aleph_i, I(j) \leq \mu_{low}(i) \text{ or } I(j) \geq \mu_{high}(i) \right\} \tag{4.19}$$

Then $|S_i^t|$ is the cardinality of the set S_i^t and $|S_i^n|$ is the cardinality of the set S_i^n, which represents the number of pixels between the high group and low group, and the number of pixels in high or low group, respectively.

To perform this bimodal test, a local threshold can be obtained for every pixel as

$$\Gamma(i) = |S_i^n| - |S_i^t| \tag{4.20}$$

The mask of the high-contrast regions, M_T, is obtained based on the threshold Γ as

$$M_T(i) = \begin{cases} 1, & \Gamma(i) > 0 \\ 0, & \text{Otherwise} \end{cases} \tag{4.21}$$

The text-like region detection method proposed in this section not only can effectively detect text-like regions in the test image but also is hyper-parameter free, which is more robust than the text-like region detection methods in Section 4.2.1 and 4.2.2.

4.3 Text Enhancement

In the Section 2.2, a band-limited local Wiener deconvolution filter [22] is introduced. The filter uses a low-pass filter B of cutoff frequency f_c to suppress the high frequency components. However, an appropriate cutoff frequency f_{c_T} for text-like regions will over-sharpen the moving detailed patterns, and a good cutoff frequency f_{c_M} for moving regions will under-sharpen text-like regions. In the proposed scheme, unlike the work in [22], we use two Wiener deconvolution kernels to enhance the input image instead of using only one kernel. Let $\tilde{g}_T$ and $\tilde{g}_\Omega$ computed in Equation (2.7) denote Wiener deconvolution kernels corresponding to the text-like regions and background, respectively. We design the two kernels using two different cutoff frequencies f_{c_T} and f_{c_Ω}, where $f_{c_T} > f_{c_\Omega}$, in order to enhance the input image I_u with different strength. Given an original image I, two enhanced images I_T and I_Ω are computed in Equation (6.1).

Then, as discussed in Section 3.3, in order to reduce the obvious boundary between more sharpened text-regions and less sharpened moving regions, a non-stationary filtering method is required to smooth the boundary. The text mask M_T obtained in Section 4.2 acting as weights of enhanced images is smoothed by the matrix w_T resulting in the weighted text mask α_T as computed in Equation (6.5), and accordingly, the weighted background mask α_Ω can be obtained in Equation (6.4) such that $\alpha_T + \alpha_\Omega = 1$. Then, the

enhanced images I_T and I_Ω are combined according to the weighted text and background mask α_T and α_Ω.

Both qualitative and quantitative evaluation of proposed text enhancement method including generated text mask will be shown in Section 7.2. Evaluation metrics (SSIM, PSNR and MSE) show the effectiveness of the proposed text enhancement method.

Chapter 5

Projector-based Motion Enhancement

In Chapter 3 the motion enhancement problem has been formulated. Current state-of-the-art resolution enhancement methods work well for static regions in a video, however, introduce temporal motion artifacts in moving regions. Therefore, moving regions should be enhanced differently from other regions with less sharpened enhancement strength based on provided motion information to avoid the artifacts. This chapter provides two solutions to enhance moving and non-moving regions with different sharpening strengths.

Section 5.1 introduces the Optical Flow-based motion enhancement scheme [2]. The pixel-wise motion velocities between two consecutive frames are obtained by proposing an improved optical flow method in which three assumptions [47] deal with larger displacement and the Kalman Filter is introduced afterwards which contributes to a better performance of motion estimation. Then a scene cut detection method is introduced to deal with the unstable motion estimation caused by shot transitions in videos. Directional blurring filters are used to blur the pixels based on their motion directions.

Section 5.2 provides a more effective and efficient video enhancement scheme which introduces a hypothesis testing-based motion detection method and, instead of blurring, enhances the moving regions with a less sharp Wiener deconvolution kernel.

5.1 Optical Flow-Based Motion Enhancement

Our goal is the development of a space-time motion estimator for the purpose of anti-aliasing (blurring) of Wiener filtering when applied to wobulated (shifted superposition) video 2.1.

Directionally localized anti-aliasing (DLAA) was developed in [48, 49] to produce anti-aliased vertical and horizontal edges by applying vertical and horizontal blurring on the image separately. Inspired by the satisfactory result of DLAA, our goal is to design a multi-directional blurring filter to blur moving regions on the basis of their motion direction.

There are many Optical Flow-based motion estimation works achieve state-of-the-art performance [47, 50, 51, 52, 53]. However, they fail to give accurage results for challenging videos with aliasing artifacts. To reduce the uncertainty of estimates a classic state estimation technique — the Kalman filter [54, 55] — can be used. Broida first proposed a recursive 3-D motion estimation algorithm [56] to estimate both the structure and kinematics of a moving object. Nikolaos and Khosla [57] proposed a real-time visual Kalman filter tracking method to track a moving object in a 2-D space. Additionally Stergios and George [58] presented an approach to simultaneously localize a group of mobile robots. These applications work well in tracking an object or a pixel in a video frame, successfully smoothing the estimated motion and reducing the error, however in our work we need a dense motion estimate for the entire image.

The Kalman filter provides a robust mechanism for image flow estimation. Kuo et al. [59, 60] improved the conventional block-matching algorithm using the Kalman filter to obtain higher precision. Singh [61] recovered image-flow from image sequences using a correlation-based approach. Cooper [62] presented an optical flow operator combined with a Kalman filter that integrates flow information across the scene to obtain two constraints on optical flow. Nonetheless, they do not consider video scene cuts, which lead to a significant error at the scene transition. In this section, we propose to use the Kalman filter to fuse motion estimates over time based on covariance information over time and scene cuts.

5.1.1 Motion Estimation using Optical Flow

An enormous amount of research literature has been dedicated to the inference and quantification of motion in video [63, 64, 65], whether for video coding, compression, tracking, or analysis. Optical flow [66] in particular is the projection of the physical movement of points to the pixel displacement on the image plane. The Combined Local-global (CLG)

Method [52] forms the basis for our work. In particular, there are three widely-asserted assumptions:

1. *Grey value constancy*: That the brightness of a pixel remains consistent [67]:

$$I(x, y, t) = I(x + u, y + v, t + 1) \tag{5.1}$$

   implying the invariant

$$I_x u + I_y v + I_t = 0 \tag{5.2}$$

   which leads to the following energy or constraint function:

$$E_{data} = \int (I_x u + I_y v + I_t)^2 dx\, dy \tag{5.3}$$

   The energy function of (5.3) can be minimized [51] as a linear system.

2. *Gradient constancy*: The gradient of the image brightness is assumed not to change because of the displacement [47], thus

$$\nabla I(x, y, t) = \nabla I(x + u, y + v, t + 1) \tag{5.4}$$

   where $\nabla = (\partial x, \partial y)$ is the spatial gradient, leading to the corresponding energy function

$$E_G = \int \left|\nabla I(x + u, y + v, t + 1) - \nabla I(x, y, t)\right|^2 dx\, dy \tag{5.5}$$

3. *Smoothness*: Horn & Schunck [68] assumes the motion field to vary smoothly, and therefore asserts an additional smoothness constraint:

$$E_{Smooth} = \sum (u_x^2 + u_y^2 + v_x^2 + v_y^2) \tag{5.6}$$

The total energy [50] is the weighted sum of these assumptions:

$$E(u, v) = E_{Data} + \alpha E_G + \gamma E_{Smooth} \tag{5.7}$$

with regularization parameters $\alpha, \gamma > 0$. By minimizing the energy term, u and v can be determined.

The problem is addressed as a global minimization leading to densely estimated motion fields [53]. However with each frame estimated independently, there is significant scope to improve the motion estimates by constraining over time, so temporal filtering, such as a Kalman filter, is desired to improve accuracy.

5.1.2 Kalman Filter

The Kalman filter is a recursive linear estimator finding optimal least-squares estimates from noisy data [69, 70, 71]. With $\hat{x}$ representing the estimate of state x, and P the estimation error covariance, the filter consists of two steps:

1. Predicting over time, based on a dynamic model F:

$$\begin{matrix} \hat{x}(t-1|t-1) \\ P(t-1|t-1) \end{matrix} \xrightarrow{F} \begin{matrix} \hat{x}(t|t-1) \\ P(t|t-1) \end{matrix} \tag{5.8}$$

2. Updating at a point in time, based on a measurement model H:

$$\begin{matrix} \hat{x}(t|t-1) \\ P(t|t-1) \end{matrix} \xrightarrow{H} \begin{matrix} \hat{x}(t|t) \\ P(t|t) \end{matrix} \tag{5.9}$$

5.1.3 Advanced Motion Estimation using Optical Flow and Kalman Filter

In our motion estimation method, the state is the motion vectors of all the pixels in each frame in the vertical direction and horizontal direction:

$$x = [u \quad v]^T \tag{5.10}$$

where u and v are the horizontal and vertical motions between two consecutive image frames separately.

In the prediction step at time t, the estimated current state $\hat{x}(t|t-1)$ are obtained by passing the estimates of the last state $\hat{x}(t-1|t-1)$ through a transition matrix F:

$$\hat{x}(t|t-1) = F\hat{x}(t-1|t-1) + w(t) \tag{5.11}$$

where F transits each pixel from previous location to the current location based on the motion information given by the previous state $\hat{x}(t-1|t-1)$ and $w(t)$ is the zero mean multivariate Gaussian noise. Each pixel is predicted to move to the new location with the same velocity according to its original location and the motion estimates of the previous state.

In the update step at time t, new information is added to the current state estimates $\hat{x}(t|t)$:

$$\hat{x}(t|t) = \hat{x}(t|t-1) + K(t)\big(z(t) - H\hat{x}(t|t-1)\big) \tag{5.12}$$

This updates the state using the newly measured motion $z(t)$ of the current frame, where $z(t)$ is obtained by in Section 5.1.1. Process noise covariance $Q = E\left[w(t)w(t)^T\right]$.

By using the Kalman Filter, some inaccurate motion velocities obtained from Section 5.1.1 are corrected. However, in practical projection, the source video always contain more than one scene and includes many scene cuts (shot transition). The estimated motion velocities for the two frames before and after a scene cut have no meaning since they are not real motion; moreover, the estimated motion velocities always indicate large motion due to the abrupt scene cut, which will cause artifacts if blurring the frame before a scene cut based on the estimated motion velocities.

5.1.4 Scene Cut Detection

In order to avoid the problem due to applying the scene-cut motion velocities as discussed in the previous section 5.1.2, a scene cut detection method is introduced in this section. Edge change ratio (ECR) [72] is a method of scoring in shot transition detection (scene cut detection). It compares the actual content between two frames in three steps:

1. Detect edges in two contiguous frames;
2. Count the number of edge pixels for each frame;
3. Determine the entering edge pixels ρ_{in} and exiting edge pixels ρ_{out}.

Then for each frame i,

$$ECR_i = \max\left(\frac{\rho_{in}}{s_i}, \frac{\rho_{out}}{s_{i+1}}\right) \tag{5.13}$$

in which s_i is the number of edge pixels in frame i.

In the decision for scene cut, we constrain the peak features to be local maxima and require the ECR to drop by 30% before considering another cut.

Fig. 5.1 shows the result of scene cut detection for the test video. The test video content is a model showing her dress.

Tests on sample videos suggest the effectiveness of ECR for scene cut detection. Then, when a frame is detected to be followed by a scene cut, the estimation of motion vectors for the current state will abandon information from the previous state; otherwise, the state can be predicted via transitions from last state.

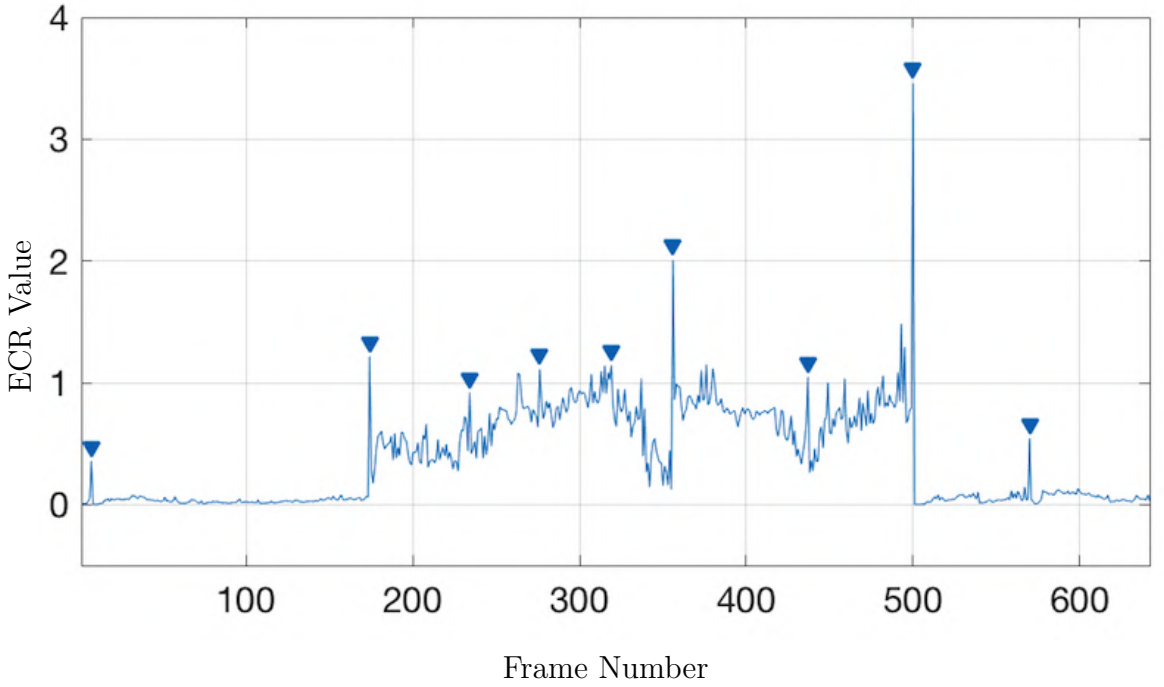


Figure 5.1: Result of scene cut detection for the test video. The higher the ECR, the more likely that a scene cut has taken place. For this test video, actual scene cuts occur at frames 7, 174, 234, 276, 321, 356, 437, 500, 570.

5.1.5 Directional Blurring Filter

Gaussian directional blurring filters are used to eliminate the artifacts either caused by SSPOS or by natural motion. Gaussian blur is a low pass filter which has the effect of reducing the image's high-frequency components. When we apply Gaussian blur on the moving regions, the high-frequency patterns in the regions can be removed, which smooths jagged edges, interference patterns and other false details aliasing in videos. Since aliasing only appear in moving areas and the simplest way to reduce aliasing in post-production is to decrease resolution by blurring the image, only blurring the moving regions while retaining the resolution of the still regions is a good tradeoff between artifacts reduction and high-resolution enhancement. We used the proposed motion estimation method in Section 5.1.3 to detect the moving regions and apply 12-directional Gaussian filters to the moving regions according to the corresponding moving directions to reduce the artifacts appearing during movement. Experimental results shown in Section 7.3.1 show that the Kalman filter can improve the motion estimation accuracy and the directional blurring filter can reduce the artifacts pattern in the test video.

5.2 Hypothesis Testing-Based Motion Enhancement

Section 5.1 introduced an Optical Flow-based motion enhancement method, which first obtain motion velocities and then blur the moving regions based on the moving direction. Though the experimental results show its effectiveness in reducing the artifacts, the Optical Flow-based motion enhancement method needs to solve the energy minimization problem, which has high computational cost [51]. In this section, we introduce a novel hypothesis-testing-based motion enhancement method. First, a motion detection method is proposed to classify each pixel into a moving or non-moving class by using statistics of the local mean-squared temporal difference. Second, a motion enhancement is proposed by sharpening moving and non-moving regions differently.

5.2.1 System Model

The overview of the proposed motion enhancement scheme is shown in the block diagram in Figure 5.2. Similar to the text enhancement scheme proposed in Figure 4.1, the text-like detection and the Wiener deconvolution for text are replaced with motion detection and Wiener deconvolution 2.2 for motion.

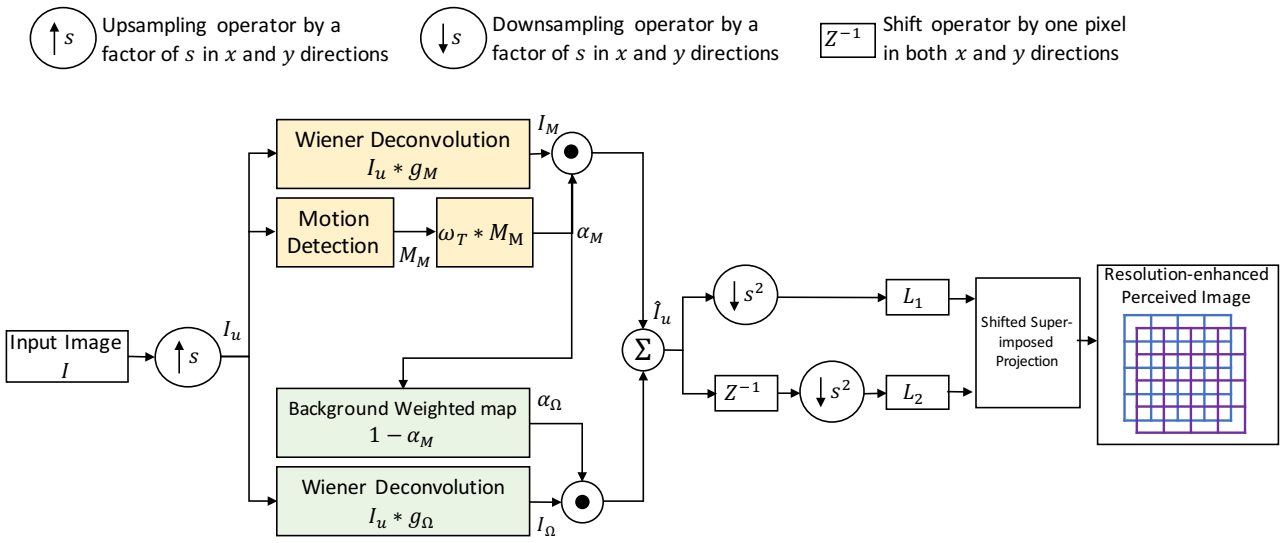


Figure 5.2: Proposed moving content enhancement scheme, where s is an upsampling factor in both x and y directions, and L_1 and L_2 are two downsampled sub-images generated with and without one pixel shift, respectively. This scheme is building on the text-like content enhancement scheme proposed in Figure 5.2.

As discussed in Section 3.2, the moving regions should be enhanced with a less sharper strength. First, an effective motion detector indicates the location of moving regions. Then, the Wiener deconvolution filter, introduced in Section 2.2, can adjust its sharpening strength by changing the cutoff frequency f_c of a low-pass filter as in Equation 2.6. Thus, a Wiener deconvolution filter with less strength is used to enhance moving regions based on the motion detector, and with sharper strength for background.

Finally, the two enhanced images are combined together and go into shifted superimposition 2.1 to generate two sub-images for projector display.

5.2.2 Motion Detection

This section states the proposed motion detection method in detail. In Section 3.2, Figure 3.5 is used to illustrate in high-level how the motion detection component in the proposed motion enhancement scheme works. In this section, Figure 5.3, as a more detailed version, describes each step of the proposed motion detection method with notations. In order to distinguish moving pixels from non-moving ones, we formulate this classification problem

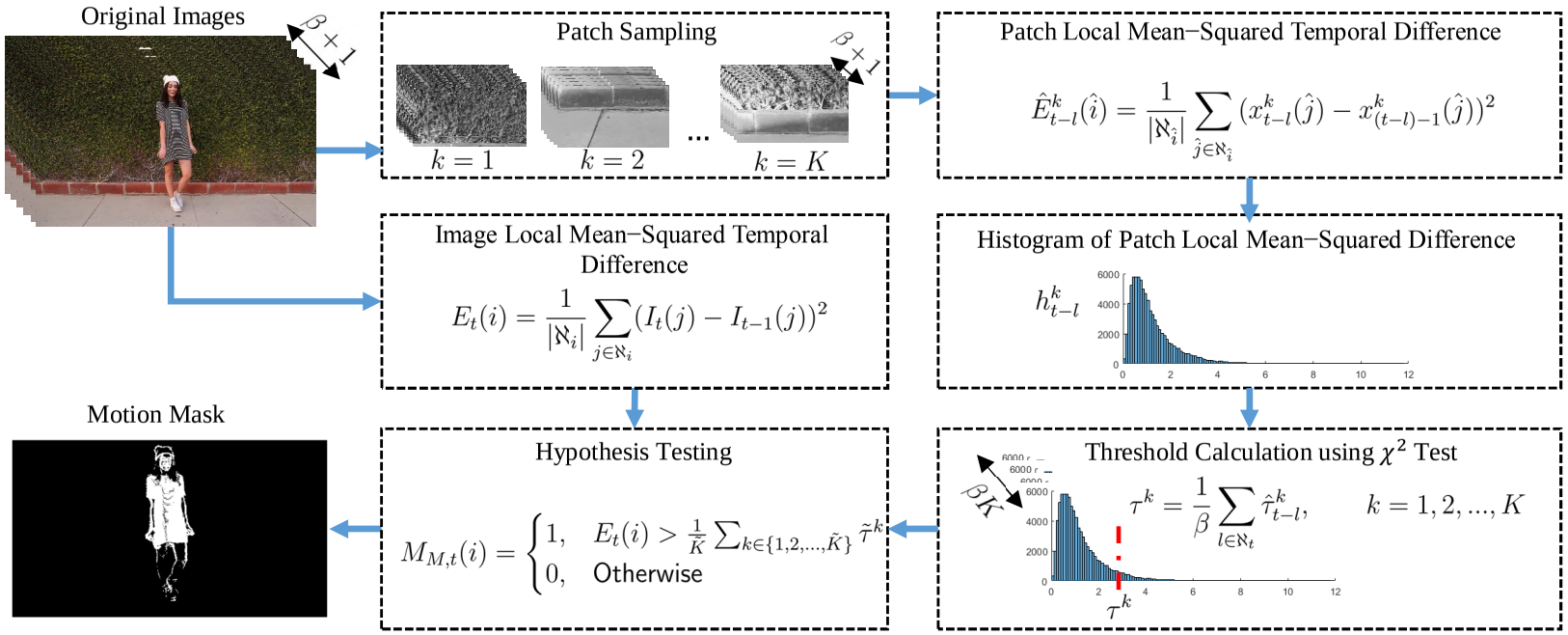


Figure 5.3: Block diagram of the proposed motion detection scheme. This figure is a detailed version of Figure 3.5. This hypothesis-testing-based motion detection method classifies each pixel into moving or non-moving class by using statistics of the local mean-squared temporal difference.

into a statistical hypothesis testing one. In the proposed scheme, a given pixel is assumed to be stationary unless an enough evidence is obtained to argue that this pixel is moving, therefore we define the null and alternative hypotheses as

- H_0: A pixel is *stationary*;
- H_1: A pixel is *moving*.

Let I_t denote the input image at a given time t. The local mean-squared temporal difference, $E_t(i)$, at time t and location i can be obtained as

$$E_t(i) = \frac{1}{|\aleph_i|} \sum_{j \in \aleph_i} \left(I_t\left(j\right) - I_{t-1}\left(j\right) \right)^2 \tag{5.14}$$

for a spatial neighbourhood $\aleph_i$ around the i^{th} pixel. Since even still video is not perfectly constant (due to camera vibration, sampling error and pixel Edge change ratio (ECR) noise) we essentially having a χ^2 problem [73]. Thus, in order to decide whether a pixel is stationary or not we need to have a threshold on E, where the threshold will need to be dynamic, as this threshold may be content–dependent.

Let the histogram of the $(k,l)^{th}$ patch local mean-squared temporal difference be denoted by h^k_{t-l},

$$\text{where} \qquad k = 1, 2, ..., K \tag{5.15}$$

$$\text{and} \qquad l \in \aleph_t = -\left\lfloor \frac{\beta}{2} \right\rfloor, -\left\lfloor \frac{\beta}{2} \right\rfloor + 1, ..., \left\lfloor \frac{\beta}{2} \right\rfloor \tag{5.16}$$

and K and β are the number of spatial and temporal sampled patches, respectively. Then, the motion threshold of the k^{th} volume location is obtained as

$$\tau^k = \frac{1}{\beta} \sum_{l \in \aleph_t} \hat{\tau}^k_{t-l}, \qquad k = 1, 2, ..., K \tag{5.17}$$

$$\text{where} \qquad \hat{\tau}^k_{t-l} = \underset{e}{\operatorname{argmin}} \left\{ e | \text{CDF}(h^k_{t-l}) \geq p, e \in E_t \right\} \tag{5.18}$$

and $\text{CDF}(h^k_{t-l})$ is the cumulative distribution function of h^k_{t-l}, p is chosen to be 0.95, and the total number of thresholds, $\hat{\tau}^k_{t-l}$, is $K \times \beta$. To minimize the effect of moving regions on computing the motion threshold, the lowest $\tilde{K} \leq K$ motion thresholds, $\tilde{\tau}^k$, are selected. Finally, the mask of moving pixels, $M_{M,t}$, is obtained by thresholding $E_t(i)$ as

$$M_{M,t}(i) = \begin{cases} 1, & E_t(i) > \frac{1}{\tilde{K}} \sum_{k \in \{1,2,...,\tilde{K}\}} \tilde{\tau}^k \\ 0, & \text{Otherwise} \end{cases} \tag{5.19}$$

5.2.3 Motion Enhancement

As discussed in Section 3.2, the enhancement filter 2.2 can not work for both moving and non-moving regions. Therefore, one possible solution is to find a new enhancement filter for both contents. However, how the filter treat each content remains uncertain. Another possible solution is to use a combination of two different Wiener filters, one for enhancing moving regions and one for enhancing non-moving regions, which provides more flexibility in designing each enhancement filter for corresponding content.

Therefore, in the proposed motion enhancement scheme, we use two Wiener deconvolution kernels to allow for content-dependent input image enhancement, instead of using only one kernel as in [22]. Let $\tilde{g}_M$ and $\tilde{g}_\Omega$ denote the Wiener deconvolution kernels corresponding to the moving and non-moving regions, respectively. We design the two kernels using two different cutoff frequencies f_M and f_Ω, where $f_M \leq f_\Omega$. Then, in order to enhance the input image I based on moving and non-moving contents, the two Wiener deconvolution kernels are applied to I respectively, and then two enhanced images are summed together according to the motion mask obtained in (5.19).

Both qualitative and quantitative evaluation of proposed motion enhancement method including generated motion mask will be shown in Section 7.3.2. Performance metrics (SSIM [74], PSNR and MSE) show that the proposed text enhancement method works better than state-of-the-art methods.

Chapter 6

Content-adaptive Resolution Enhancement

Section 3.3 formulates the content-adaptive enhancement problem. Since the regions with different contents are enhanced differently in Chapter 4 and Chapter 5, if directly combining enhanced text-like regions, moving regions and background together, there will be a sharp boundary between these three contents especially when moving and text-like regions are adjacent. Most existing projector-camera systems do not enhance the resolution in a content-adaptive way, and none of them have considered the problem of combining contents with different features. This Chapter provides the solution of generating a final composite image by smoothly combining moving regions, text-like regions and background, which is implied by the summation symbol in Figure 6.1.

6.1 Introduction

In order to solve the gap between enhanced text-like regions, moving regions and background, the simplest idea is to smooth the gap regions based on a text-like mask M_T obtained in Equation 4.21, a motion mask M_M obtained in Equation 5.19 and a background mask M_Ω. However, smoothing the gap regions will also degrade the image quality. Therefore, we propose to apply non-stationary filtering to the three enhanced images I_T, I_M, and I_Ω with different sharpening levels. The idea is to assign a weight factor to each enhanced image when combining them.

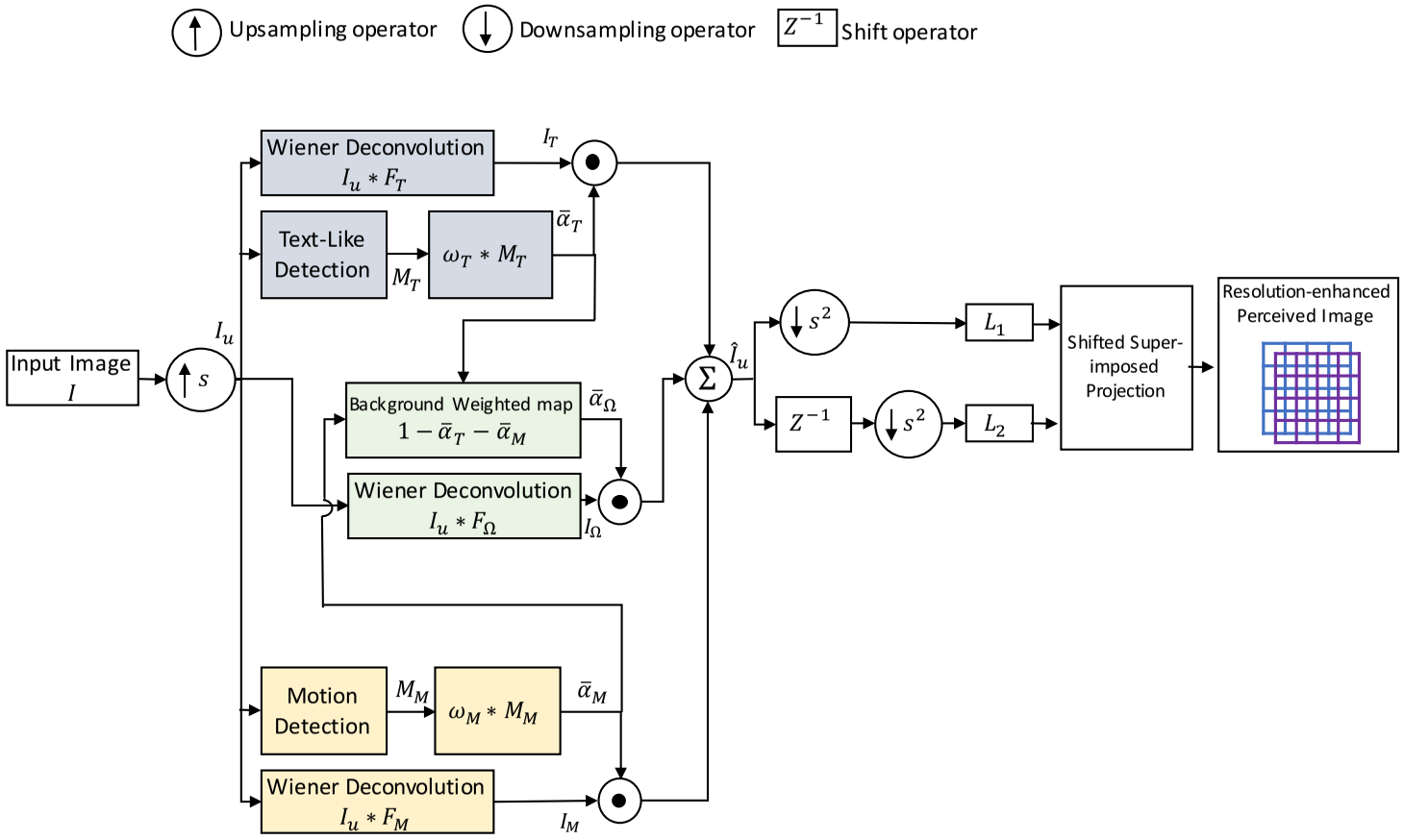


Figure 6.1: The proposed non-stationary content-adaptive enhancement scheme. This figure has been shown in Figure 3.6 to give an overview of content-dependent enhancement for text (grey), motion (yellow) and background (green) using different enhancement kernels. Further details are explained in Chapter 6. Each component can lead to the improvement of perceptual quality of images based on the feature of the contents. The content detection methods, the selection of the corresponding enhancement kernel and the fusion of enhanced images are the focus in this work.

6.2 Non-Stationary Filtering

First, given an original image I, three enhanced images I_T, I_M, and I_Ω with different sharpening levels are generated. Define a set $G = \{T, M, \Omega\}$, where T, M and Ω stand for text, motion and background, respectively. I_T, I_M, and I_Ω are obtained by applying three Wiener filters F_T, F_M and F_Ω with different sharpening parameters to I respectively:

$$I_k(i,t) = I * F_k \quad \forall k \in G \tag{6.1}$$

After obtaining the three enhanced images I_T, I_M, and I_Ω, in order to generate a composite image with a smooth transition between different contents, I_T, I_M, and I_Ω should be summed based on three corresponding weight factors. For instance, the weight factor for I_T should has higher value in text-like regions, zero value in non-text-like regions, and the value in between in gap regions. So the goal now is to find a set of local weights $\bar{\alpha}_T$, $\bar{\alpha}_M$, and $\bar{\alpha}_\Omega$ that compute the composite image such that we have more sharpened text in I_T, less-sharpened motion in I_M, and enhanced background in I_Ω all contribute to an enhanced composite image $\hat{I}$ as follows:

$$\hat{I}(i) = \sum_{k \in G} \bar{\alpha}_k(i,t) \circ I_k(i,t) \tag{6.2}$$

Let M_T, M_M, and M_Ω denote a text-like mask obtained in Equation 4.21, a motion mask obtained in Equation 5.19 and a background mask. Since all of the masks M_T, M_M, and M_Ω have a value equals to one on related-content regions and zero on other regions, these masks can be used to calculate the weight factor $\bar{\alpha}_k$. First, each mask needs to be smoothed in order to make the pixel value of the smoothed mask in transition regions between zero and one:

$$\alpha_k(i,t) = \begin{cases} \sum_{l=t} \sum_{j \in \aleph_i} \bar{w}_k(i,j,t,l) M_k(i,j,t,l) & \text{if } k = T \\ \sum_{l \in \aleph_t} \sum_{j \in \aleph_i} \bar{w}_k(i,j,t,l) M_k(i,j,t,l) & \text{if } k = M \end{cases} \tag{6.3}$$

$$\alpha_\Omega(i,t) = 1 - \sum_{k \in G \setminus \{\Omega\}} \alpha_k(i,t) \tag{6.4}$$

Second, to keep the pixel values of composite image in the same range of the that of the sharpened image I_k, the weight factor $\bar{\alpha}_k$ is obtained by normalizing α_k

$$\bar{\alpha}_k(i,t) = \frac{\alpha_k(i,t)}{\sum_{g \in G \setminus \{\Omega\}} \alpha_g(i,t)} \tag{6.5}$$

$$\ni \quad \sum_{k \in G} \bar{\alpha}_k(i,t) = 1$$

The weight factor $\bar{\alpha}_T(i,t)$ uses a text map M_T of the original image as well as a normalized penalty weight $\bar{w}_T$ to determine the impact of each pixel i of image I_T at time t within spatial neighbourhood $\aleph_i$, to the final composite image $\hat{I}$. Moreover, $\bar{\alpha}_M(i,t)$ is the normalized weight that corresponds to the less-sharpened motion image I_M. $\bar{\alpha}_M(i,t)$ uses a motion map M_M of the original image as well as a penalty weight $\bar{w}_M$ to determine how each pixel i of image I_M at time t within both spatial neighbourhood $\aleph_i$ and temporal neighbourhood $\aleph_t$ contributes to the final composite image $\hat{I}$. In a similar fashion, $\bar{\alpha}_\Omega$ for background is computed in (6.4).

In order to make the normalized penalty weights $\bar{w}_k(i,j,t,l)$ to be a smoothing filter spatially and temporally, $\bar{w}_k(i,j,t,l)$ are calculated based on the Euclidean distance $d_E(\cdot)$ between pixels i and j ($j \in \aleph_i$) and temporal distance $d_{tmp}(\cdot)$ between frames at times t and l ($l \in \aleph_t$):

$$\bar{w}_k(i,j,t,l) = \begin{cases} \frac{w_T(i,j)}{\sum_{n\in\aleph_i}\left(w_T(i,n)+\sum_{u\in\aleph_t} w_M(i,n,t,u)\right)} & \text{if } k = T \\ \frac{w_M(i,j,t,l)}{\sum_{n\in\aleph_i}\left(w_T(i,n)+\sum_{u\in\aleph_t} w_M(i,n,t,u)\right)} & \text{if } k = M \end{cases} \tag{6.6}$$

where

$$w_T(i,j) = \exp\left(-\frac{d_E^2(i,j)}{\sigma_{sp}^2}\right) \tag{6.7}$$

$$w_M(i,j,t,l) = \exp\left(-\frac{d_E^2(i,j)}{\sigma_{sp}^2} - \frac{d_{tmp}^2(i,j,t,l)}{\sigma_{tmp}^2}\right) \tag{6.8}$$

with σ_{sp} and σ_{tmp} being spatial and temporal control parameters that determine how much farther pixels/frames contribute, such that:

$$\sum_{j\in\aleph_i} \bar{w}_k(i,j,t,l) = 1 \quad \mathrm{k} \in G \setminus \{\Omega\} \tag{6.9}$$

After obtaining the composite image $\hat{I}$, shifted superimposition introduced in Section 2.1 will be used to obtain two sub-images, L_1 and L_2, by shifting $\hat{I}$ by half a pixel.

6.3 Conclusion

This chapter introduces a content-adaptive projector resolution enhancement scheme that smoothly combines differently enhanced text-like and moving regions. A non-stationary

filtering is introduced to assign a weight factor to each enhanced image based on its mask. Then, an enhanced composite image is obtained by adding the weighted enhanced images, where more sharpened text, less-sharpened motion, and enhanced background all contribute to the enhanced composite image. The quantitative and qualitative results in the following Section 7.4 show that the non-stationary filtering successfully combining differently enhanced contents smoothly.

Chapter 7

Experimental Results

7.1 Datasets

A Visual Projection Assessment Dataset (VPAD) has been created which consists of a set of videos collected from various movies, documentary, sports and TV news channels. The video sequences were obtained from a wide range of websites such as [75] and [76]. The dataset includes a total of 233 video sequences and is publicly released[1] to encourage further research into projector resolution enhancement assessment in practical environments.

The dataset includes the following ten categories: Action Movies, Comedy/Romance Movies, Documentary, Fantasy Movies, Graphics or Animation, Horror Movies, News, Sports, TV Shows, TV Episodes. The videos from the same category may share some common features, such as similar background, similar content, etc. A detailed summary of this dataset is shown in Figure 7.2.

Having discussed in Chapter 3, it is known that the text-like regions, moving regions and background have distinct features. Thus, when given an input video containing a combination of these features, the projector projection quality is very likely to be affected. However, there is no public video dataset can be accessed containing text-like regions, moving regions and background simultaneously, and targeting on projector projection assessment, which makes it hard to evaluate the projector projection. Considering this, the VPAD video dataset is created with the presence of moving regions and text-like regions in videos.

[1] URL: http://vip.uwaterloo.ca

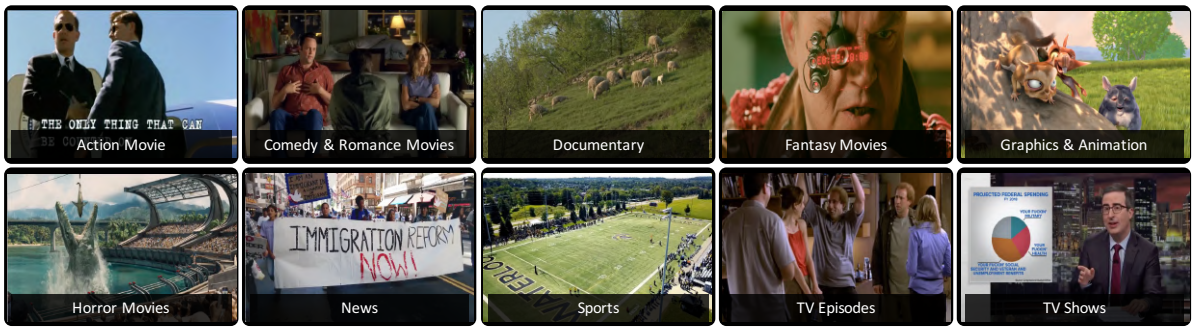


Figure 7.1: Sample Videos of the Visual Projection Assessment Dataset (VPAD). VPAD is a projection assessment video dataset aiming to provide a large variety of test videos to evaluate the quality of projector projection. VPAD includes a total of 233 sequences in 10 categories. Each video includes both text-like regions and moving regions.

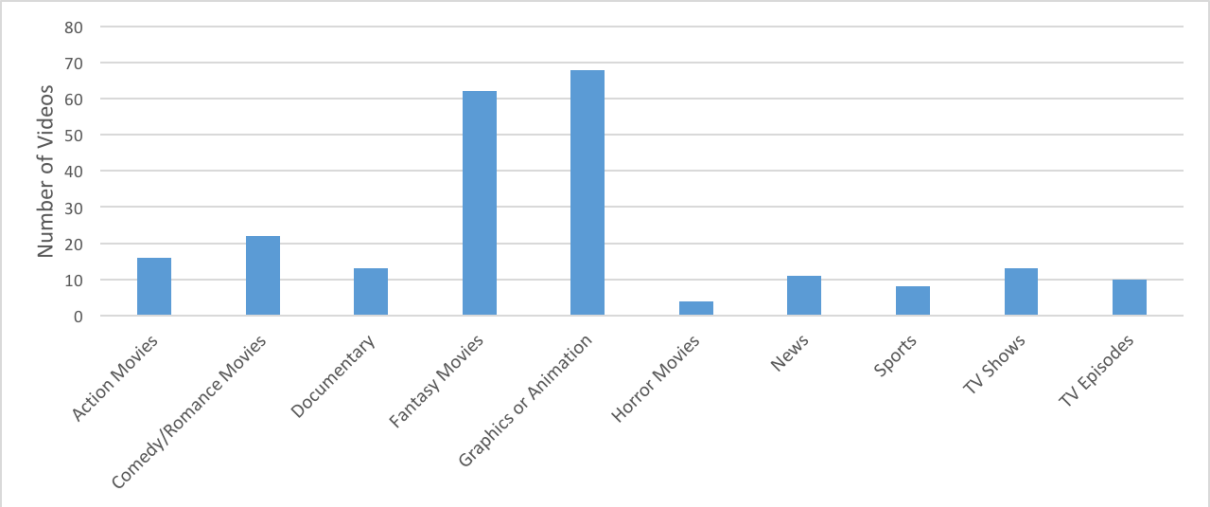


Figure 7.2: Overview of the Visual Projection Assessment Dataset (VPAD).

Table 7.1: Quantitative results of the proposed text enhancement method, Ma *et al.* [22] with different settings and projection without enhancement on the VPAD dataset, where SSIM, MSE and PSNR are used.

Method	SSIM$\times 10^{-2}$	MSE$\times 10^{-3}$	PSNR
Proposed Method (f_T:30, f_B:28)	**99.98**	**9.30**	**20.35**
Ma *et al.* ($f_c = 28$) [22]	**99.98**	<u>9.98</u>	<u>20.07</u>
Without Enhancement	**99.98**	12.9	18.95

Note: The best and the second best results on each sequence are shown in boldface and underscore, respectively.

7.2 Text Enhancement Results

The proposed text-like region enhancement method where the detected text-like regions are obtained in Equation 4.17 is evaluated on the VPAD dataset. Table 7.1 shows the quantitative comparison between the proposed text enhancement method and Ma *et al.* [22].

The proposed text-like region enhancement method has been tested on a $120Hz$ Christie projector[2] 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 7.4. The test images include different types of text-like and background regions. The *Eyechart* image has the text of different scales and background of gradually varying intensity. For the *Video Card* image, it has text regions of 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 the 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 7.4 shows the text masks of the proposed scheme on the test images into consideration. It is clear that the proposed scheme has been able to detect text-like regions of different fonts, styles and orientations. Figure 7.4 also provides a qualitative comparison among the proposed method, projection without enhancement

[2]Christie Matrix StIM WQ simulation projector

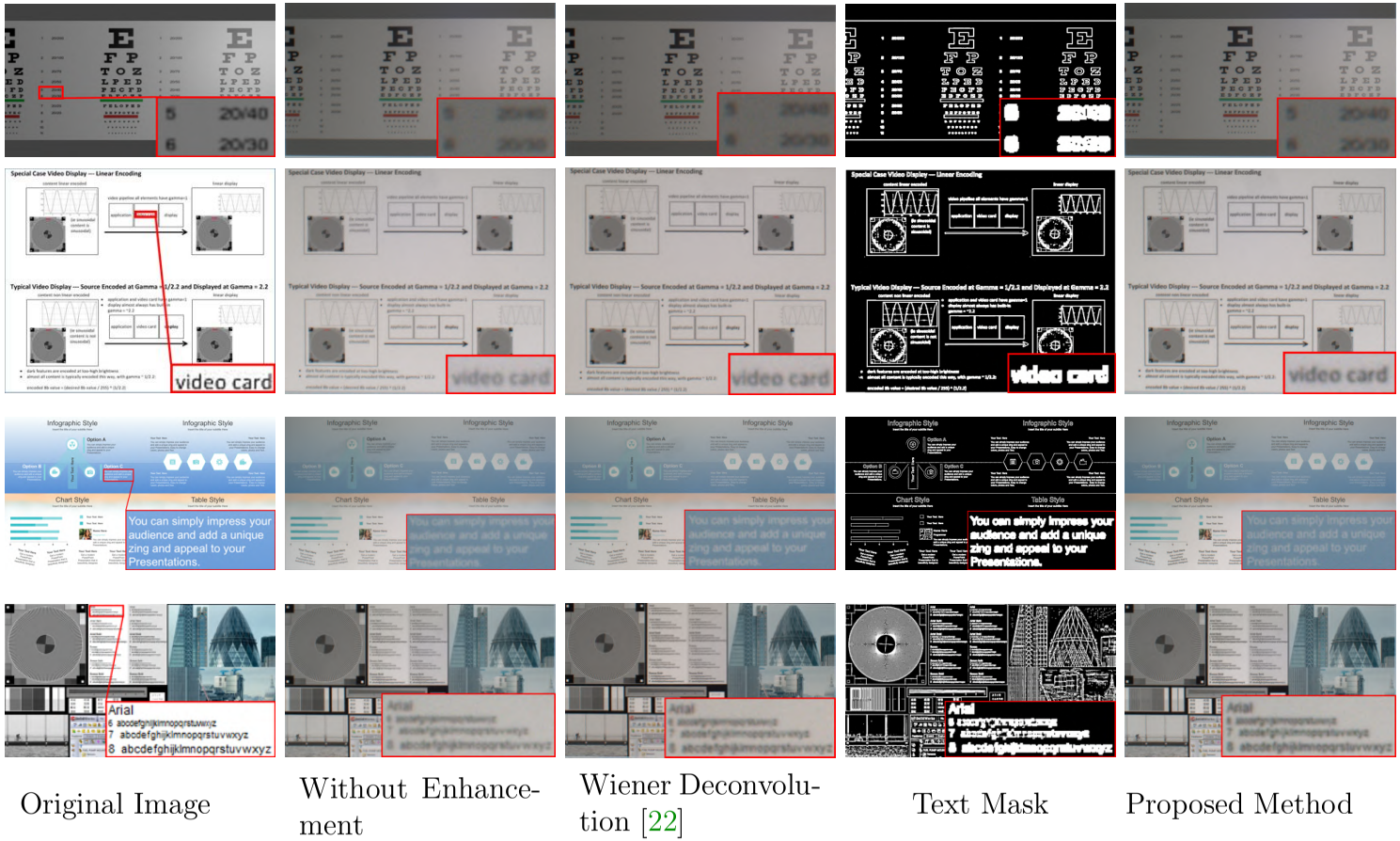


Figure 7.3: Qualitative results of the proposed method (4^{th} and 5^{th} columns), projection without enhancement (2^{th} column) and Wiener Deconvolution [22] (3^{rd} column) on four original images (1^{st} column). It is observed from the results that the proposed method offers a better visual resolution quality than that offered by the projection without enhancement and the method in [22], especially for text-like regions. For instance, the words "video card" in the second row of Figure 7.4 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 7.4, 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.

([6]) and Wiener Deconvolution method [22], where these methods have been tested on each image and the actual projection outputs were photographed by a camera from the projection screen. Our proposed method was able to sharpen text more while avoiding over sharpening background for all test images.

7.3 Motion Enhancement Results

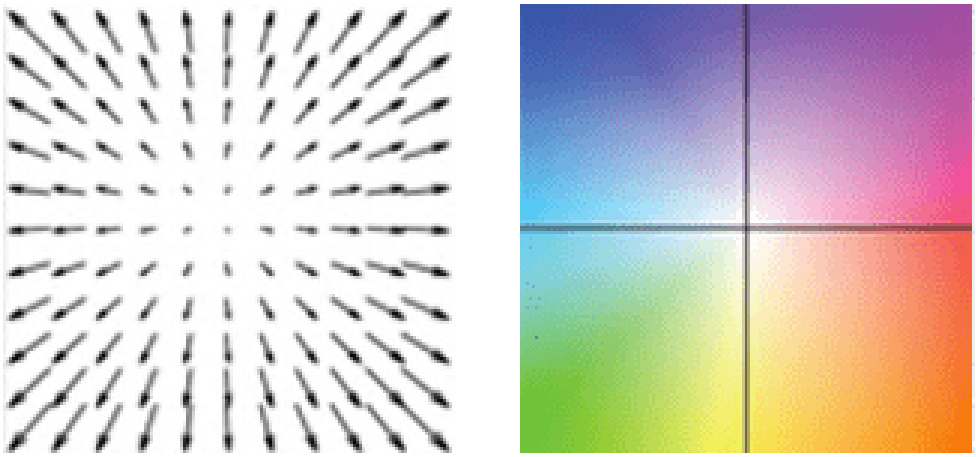
Considering that an effect filter for static regions will over-enhance the moving regions causing motion artifacts while a filter suitable for moving regions will lead to under-enhance the other regions, in this thesis, two motion enhancement schemes are proposed in Chapter 5 to deal with this problem by enhancing moving and static regions with different sharpening levels, respectively. The Optical Flow-Based motion enhancement scheme sharpens the moving regions obtained in Equation 5.12 based on the velocities of moving regions. Considering the efficiency of this method, a hypothesis testing-based motion enhancement method is proposed where the moving regions are detected in Equation 5.19.

7.3.1 Optical Flow-Based Motion Enhancement Results

While there is always noise appearing in the estimated motion, we used the Kalman filter to correct the estimated motion based on the motions of previous frames and current measurement. In this section, our proposed motion estimation approach was performed on various test videos with different scene characteristics: synthetic and real-world scenes. The simulation results show that the performance of Optical Flow-based motion detection is improved by using the Kalman Filter.

The flow visualization in Fig. 7.4 denotes the relationship between the motion map and the motion vectors. The color of the motion map denotes the direction and magnitude of the motion for each pixel.

Fig. 7.5 shows the results of our method tested on the "spinning" video and "moving-lines" video. The motion maps generated by the proposed method are cleaner than Optical Flow-based motion detection without the Kalman Filter and the other state-of-art methods. For example, the motion magnitude of the dark red square inside the red motion map of the proposed method is almost twice than the ground truth motion magnitude. After we applied the proposed method, the estimated motion became much closer to the true motion. The new flow maps of the proposed method give a qualitative impression of their motions: they are very consistent with the observed motion. Not only is the discontinuity between



Flow Fields Visualized Color

Figure 7.4: Flow visualization. Mapping flow vectors to corresponding color with different intensity according to the magnitude and direction of the velocity. We follow the code in [77] to visualize the flow field.

the different types of motion preserved, but also the translational motion is estimated. By comparing the mean square error listed in Table 7.3, we can see that the errors in the result of the previous method have been greatly weakened in the result of the proposed method. Nevertheless, the mean square error of the proposed method and previous method in Table 7.2 is similar. The reason is that the motion of the previous method is already consistent between frames while the Kalman filter works well in filtering sudden errors between frames.

We tested our algorithm in four various videos shown in Fig. 7.6. We picked the frames whose motion maps have abrupt errors using our previous method. By inspecting visually, the estimated motion became better after we used the Kalman filter because the error had been reduced. In general, our proposed result gives more accurate, dense and smooth motion flow fields and performs better in motion estimation.

In the state transition function, we considered the situation of scene cuts. The state of the current frame where a scene cut happened is not related to the state of the previous frame. If we still use the previous motion information as a predictor for the current state, the motion map will have more errors. Here the ECR algorithm we used can do a very good job in detecting scene cuts. We compared two kinds of motion maps for the frames where scene cuts happen: without or with the consideration of the the previous frame. From Fig. 7.7 we can see that the irrelevant moving information of previous frame will

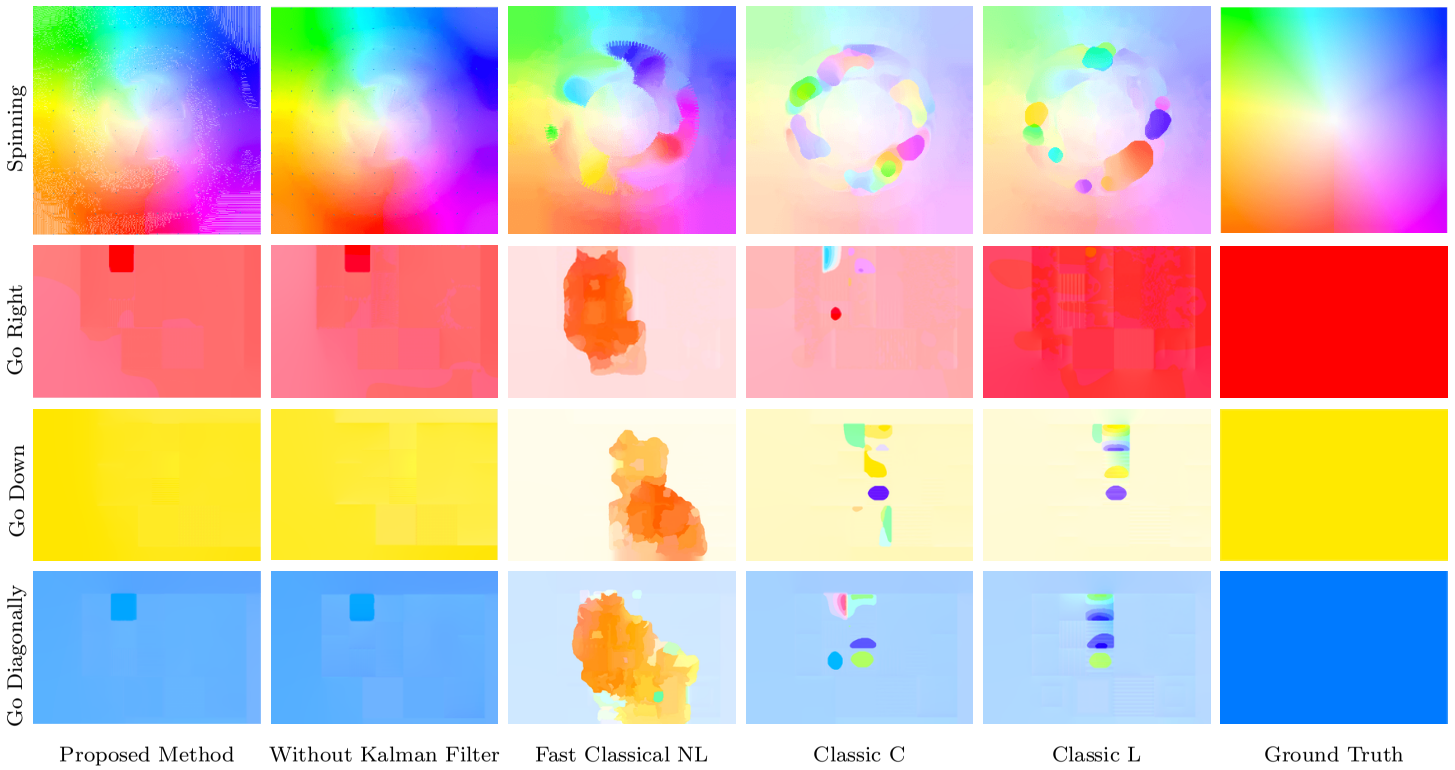


Figure 7.5: Result comparison between our proposed method, without Kalman Filter method [2], Fast Classical NL, Classic C, Classic L [78, 79]. The first row is the motion map of a anti-clockwise spinning video. The motion map of proposed method and the previous method is very similar to the true motion. For the "spinning" video, the motion map of the proposed method did the same good as the previous method because there are no sudden errors appearing in the result of the previous method. Hence the Kalman filter does not have much things to do with the motion results. For the motion map of the other three methods, the motion of the center circle is not accurate. For example, red means moving to the right. So the bottom of the circle should be red whereas the colors in the three motion maps are not correct. The second row to the last row are motion maps of the video the patterns of which are first moving to the right, downwards and diagonally from the lower right corner to the upper left corner. Our proposed method has increased the accuracy of previous method and has fewer errors than the others.

Table 7.2: Average mean square error for synthetic test case: "spinning" video. The proposed method applies the Kalman Filter to Optical Flow-based motion estimation method, where motion velocities are obtained in Equation 5.12. The method without Kalman Filter is the Optical Flow-based motion estimation method estimating motion in Equation 5.7.

Method	MSE of horizontal Velocity	MSE of Vertical Velocity
Proposed Method	0.09	0.09
Without Kalman Filter	0.09	0.09
Fast Classic NL	0.22	0.23
Classic C	0.75	1.04
Classic L	0.46	0.52

Table 7.3: Average mean square error for synthetic test case: "moving lines" video

Method	MSE of horizontal Velocity	MSE of Vertical Velocity
Proposed Method	0.34	0.15
Without Kalman Filter	0.72	0.35
Fast Classic NL	94.39	92.76
Classic C	1.64	1.39
Classic L	0.41	2.81

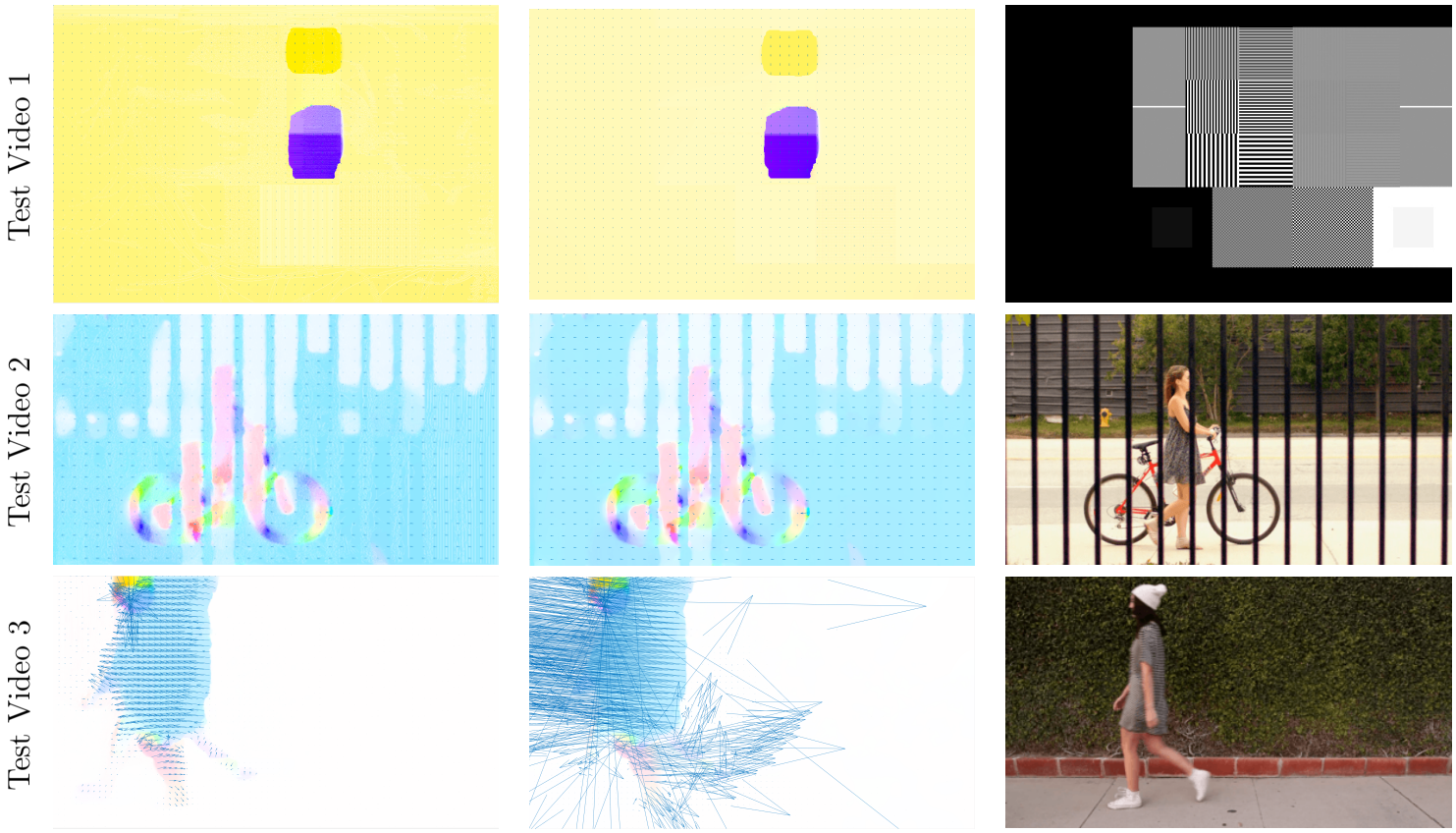


Figure 7.6: Qualitative results comparison on different videos. For the first video, the color of the center of the motion map is different than the others. The purple square in the motion map means that region is moving upwards which is opposite to the real motion. The error is suddenly appeared in this frame while the motion can be correctly estimated among its adjacent frames. Our proposed method has greatly reduced the error and let the motion values of the other regions closer to the ground truth based on the information of the previously estimated motions. The correction is more obvious in the third test video. There are many messy motion vectors in the result of previous method. And after correction, the incorrect vectors have been eliminated.

disturb the estimation of the following frame.

Fig. 7.8 shows the final result after applying adaptive Gaussian blurring kernels to moving regions based on the motion we estimated. The artifacts appearing by movement in the video are weakened by our final results.

7.3.2 Hypothesis Testing-Based Motion Enhancement Results

The proposed motion enhancement method where motion regions are detected in Equation 5.19 has been tested on a 120Hz Christie Digital projector,[3] which includes a piezo-electric actuator introducing a diagonal half-pixel shift. A software–triggered RGB camera was positioned to capture the superimposed projection results. The proposed scheme was evaluated on four videos, namely, *Spinning* (360 frames at 276×276), *RaceCar* (120 frames at 1920×1080), *Girl* (642 frames at 1920×1080) and *ToyTrain* (348 frames at 1280×720). The sample images are shown in the first column of Figure 7.10. These videos include multiple representations of moving and background regions.

Quantitative Evaluation: To evaluate the performance of the proposed scheme in enhancing projected imagery, we compare the results of our method with that of projection without enhancement and the recent method presented in [22]. For this purpose, the Structural SIMilarity (SSIM) [74], Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR) are used.

Table 7.4 shows the values of these metrics for the proposed method, projection without enhancement and the method in [22] on the *Spinning* videos. Although all the three evaluation metrics indicate that the Wiener Deconvolution approach with cut-off frequency $f = 34$ is the best method, on the contrary, it produces severe aliasing shown in the upper right image in Figure 7.9. Besides, it is seen that the SSIM score of the proposed method is very similar to the method in [22]. We conclude that SSIM, MSE and PSNR are doing poorly in evaluating the video quality with motion artifacts. Thus, the comparisons using SSIM, MSE and PSNR are all not very meaningful, and we suggest finding other metrics for assessing aliasing.

Motion Artifacts Measurement Since all of the current metrics fail in detecting aliasing introduced in high-frequency patterns and seem to prefer the methods that introduce less blur instead, no matter how severe aliasing in the moving regions is introduced, we propose a new indicator to measure the degree of aliasing artifacts in a given video based on the temporal information.

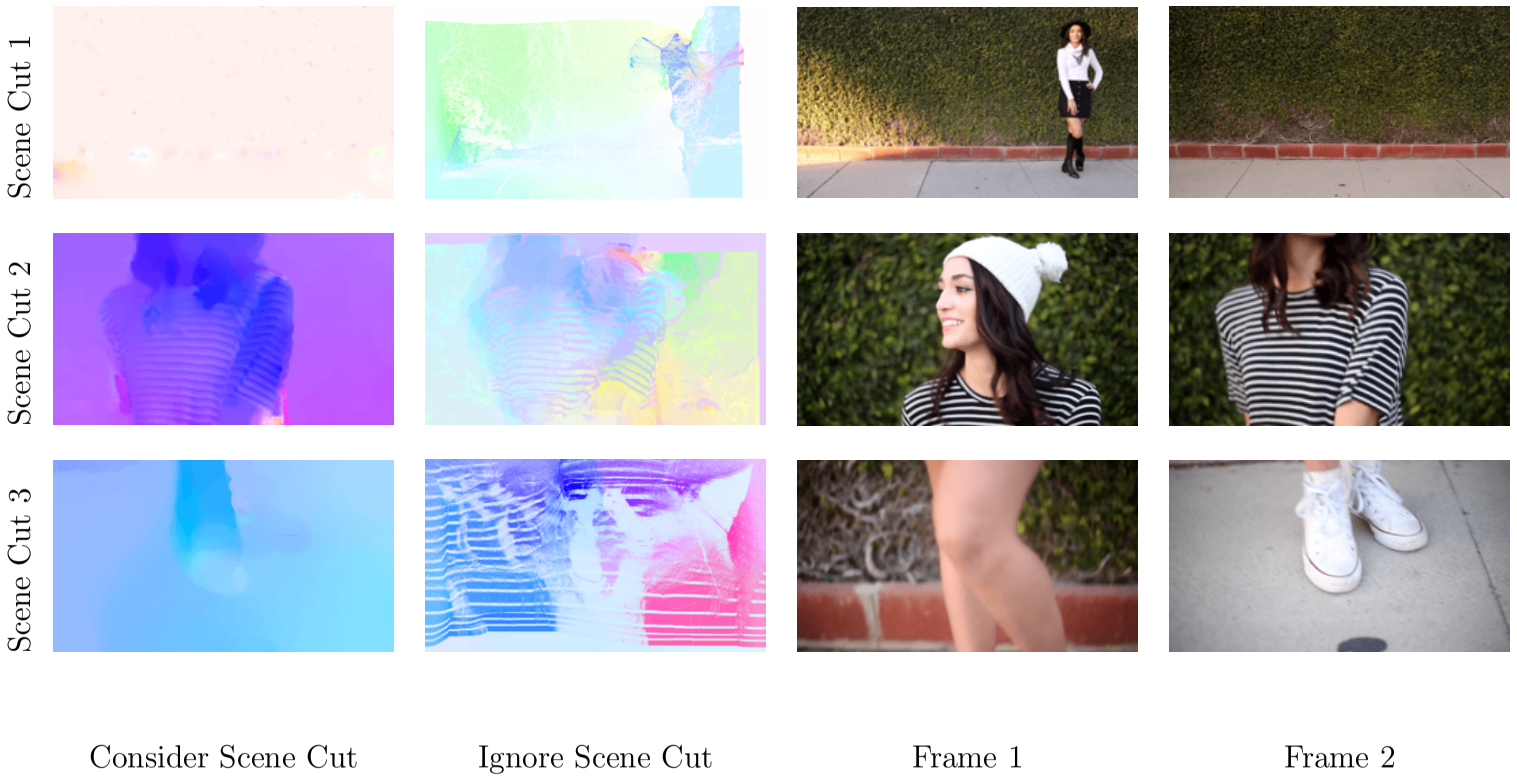
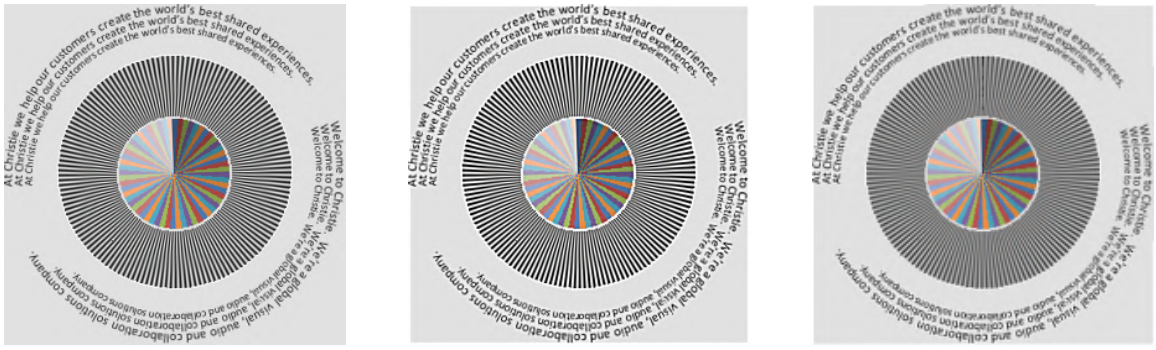


Figure 7.7: Scene cut detection applied on motion estimation. This figures shows that when there is a scene cut happened between two consecutive frames, the motion information should be abandoned when estimating the motion. The motion map that considers scene cut (motion map of two frames after scene cut) is clean and accurate whereas the motion map that ignores the scene cut is incorrect because it considers the motion information of last irrelevant scene.



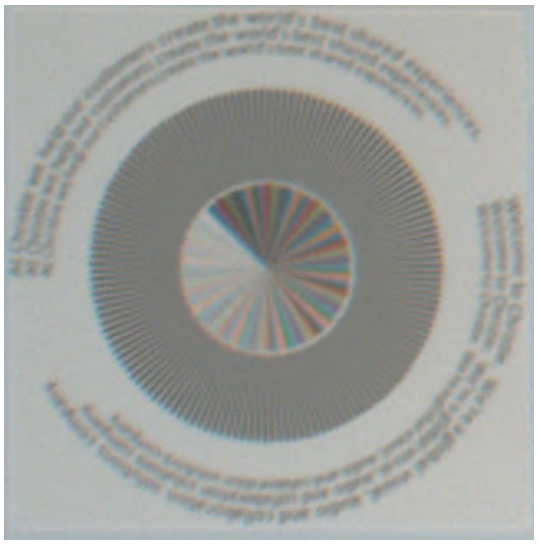
No Wiener (a) Wiener Deconvolution (b) Gaussian Blurring (c)

Figure 7.8: Qualitative results using directional blurring. (a) is the original image. We can see the artifacts inside the fringe pattern surround center. (b) is the enhanced high resolution images generated using Wiener deconvolution. The artifacts pattern on the top, bottom, left and right of the image can be more clearly observed. (c) is the image blurred by our Gaussian filters according to the motion we estimated. The blurring filter is of 12 directions. From the image we can see that the image is blurred smoothly and the artifacts have been greatly weakened in the blurred image. In the meanwhile, video resolution is nicely enhanced.

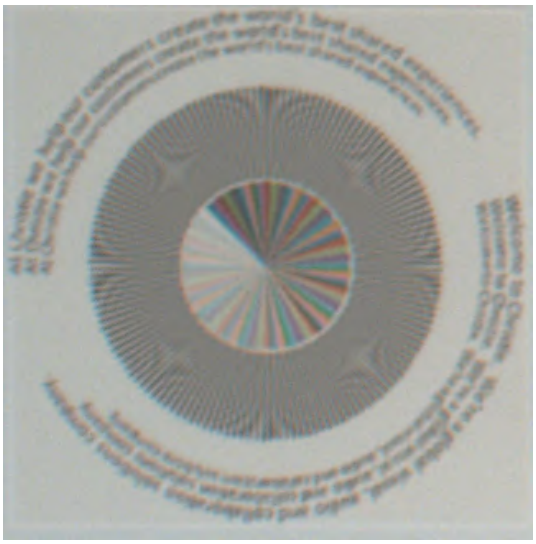
Table 7.4: Quantitative results of the proposed motion enhancement method, Ma *et al.* [22] with different settings and projection without enhancement on the *Spinning* sequence, where SSIM, MSE and PSNR are used.

Method	SSIM$\times 10^{-2}$	MSE$\times 10^{-3}$	PSNR
Proposed Method (f_M:28, f_B:32)	99.79	26.17	15.82
Proposed Method (f_M:32, f_B:34)	<u>99.84</u>	<u>19.84</u>	17.02
Ma *et al.* ($f_c = 32$) [22]	99.82	19.84	<u>17.03</u>
Ma *et al.* ($f_c = 34$) [22]	**99.86**	**17.98**	**17.45**
Without Enhancement	99.76	31.66	15.00

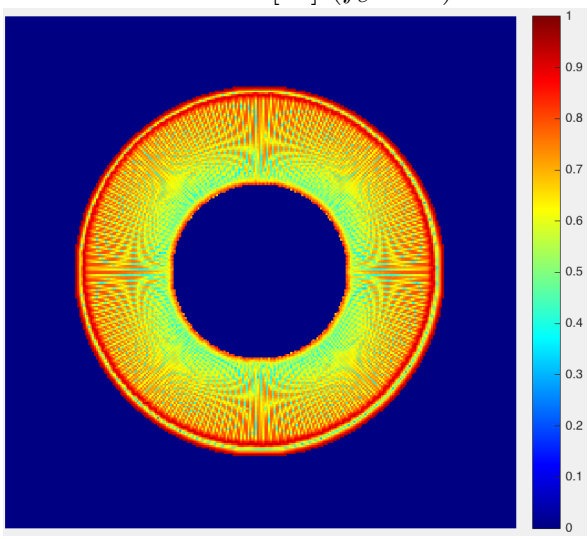
Note: The best and the second best results on each sequence are shown in boldface and underscore, respectively.



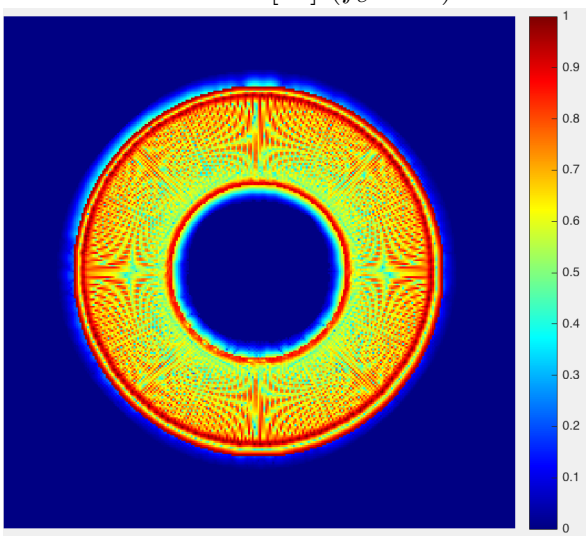
Ma *et al.* [22] ($f_c = 32$)



Ma *et al.* [22] ($f_c = 34$)



$\tilde{V}$ of proposed method ($f_M = 28, f_\Omega = 36$)



$\tilde{V}$ of Ma *et al.* [22] ($f_c = 36$)

Figure 7.9: Comparing the motion artifacts in projected images enhanced by different levels and comparing the magnitude of optical flow, $|\tilde{V}|$, as an assessment of temporal aliasing for two-parameter settings.

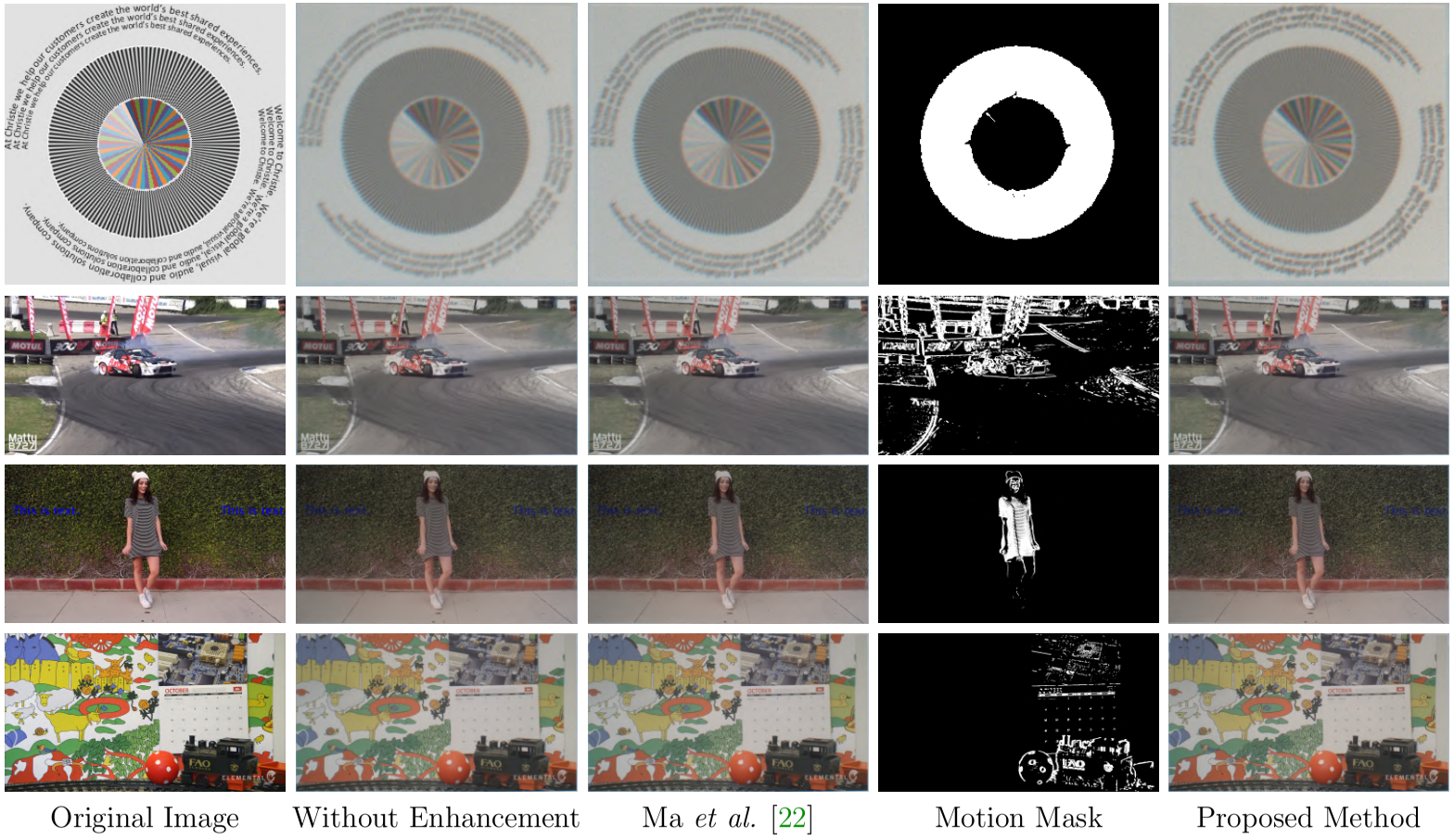


Figure 7.10: Qualitative results of the proposed method (4$^{\text{th}}$ and 5$^{\text{th}}$ columns), projection without enhancement (2$^{\text{th}}$ column) and Ma *et al.* [22] (3$^{\text{rd}}$ column) for four original images (1$^{\text{st}}$ column). Please see text for discussion.

For this purpose, optical flow [68] has been used to show how close the enhanced video is after projection to the original one. Since optical flow calculates the apparent content velocities within two successive frames, we believe that the aliasing artifacts will result in additional velocities, which can be measured numerically. To develop the new metrics, let the optical flow error $\tilde{V}_t^q$ between the true velocity $V_t^q(i)$ and estimated velocity $\hat{V}_t^q(i)$ at time t, direction q and location i be defined as

$$\tilde{V}_t^q(i) = \hat{V}_t^q(i) - V_t^q(i) \tag{7.1}$$

Then, the spatial and temporal variances of the optical flow errors are, respectively, obtained as

$$\sigma_{spa} = \frac{1}{QT} \sum_{q,t} \text{var}\left(\left\{\tilde{V}_t^q\left(i\right), i = (1,1), (1,2), ..., (N_1, N_2)\right\}\right) \tag{7.2}$$

$$\sigma_{tmp} = \frac{1}{N_1 N_2 Q} \sum_{i,q} \text{var}\left(\left\{\tilde{V}_t^q(i), t = 1, 2, ..., T\right\}\right) \tag{7.3}$$

where var(.) computes the statistical variance, N_1 and N_2 denote the numbers of pixels in the x and y directions, respectively, Q is the number of motion directions, *i.e.*, two directions, and T is the number of frames. The second row of Figure 7.9 shows the magnitude of the velocities calculated using the optical flow calculation between the 9^{th} and the 10^{th} frames of the *Spinning* video. It is noticed from this figure that the average $|\tilde{V}|$ of pixels in moving regions increases by increasing the sharpening level of the Wiener deconvolution kernel corresponding to these regions. Table 7.5 shows the motion artifacts quantified using the spatial and temporal variances, σ_{spa} and σ_{tmp}, calculated for the moving regions of the *Spinning* video. This table confirms that the proposed quantitative measures in (7.2) and (7.3) are capable to indicate the severity of the visual motion artifacts caused by applying Wiener deconvolution kernels at different sharpening levels.

Qualitative Results: Figure 7.10 shows sample qualitative results for projection without enhancement, the method in [22], and the motion mask and the enhanced image for the proposed technique on the four test videos. Evidently, the proposed scheme allows sharpening the background regions without affecting moving, resulting in superior visual quality.

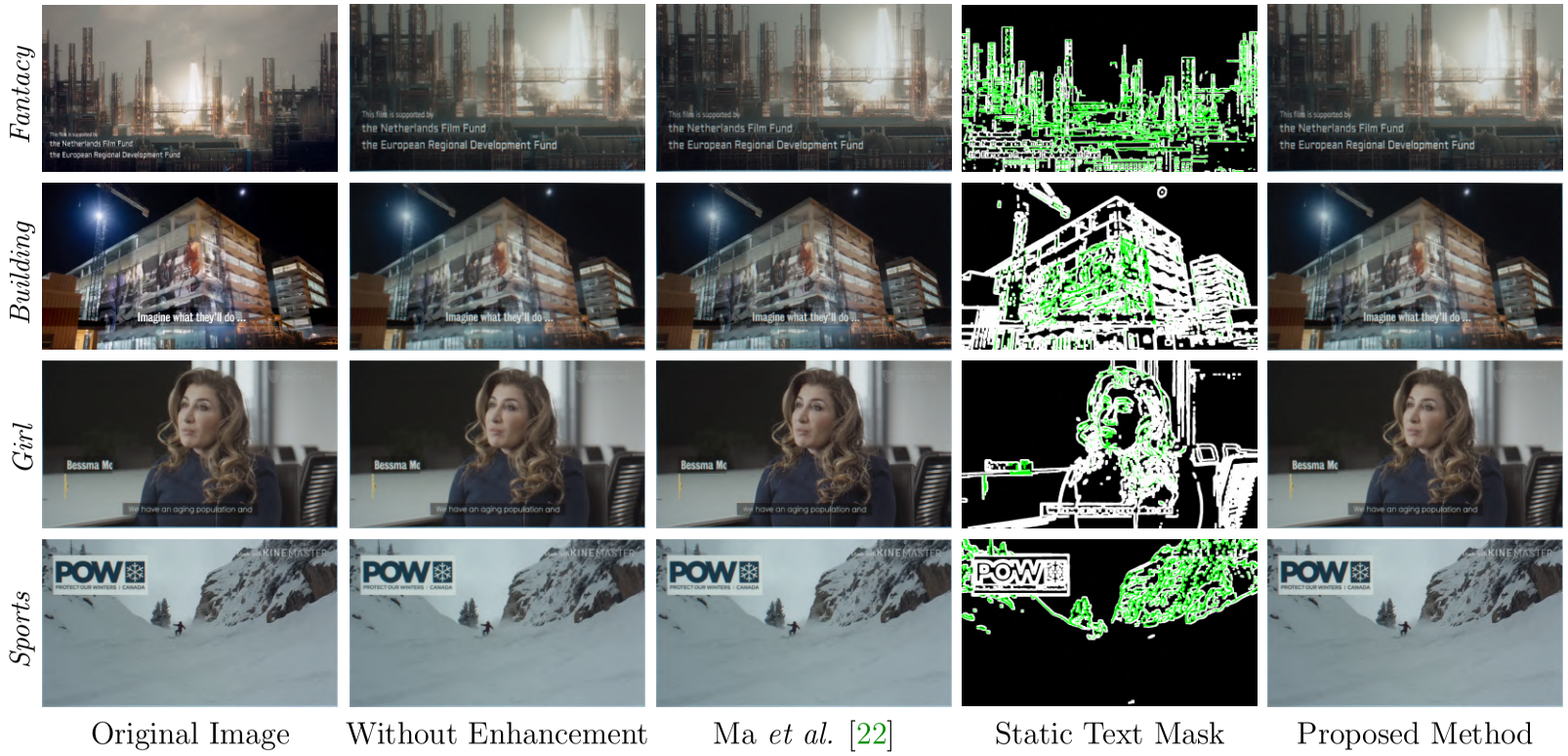


Figure 7.11: Qualitative results of the proposed non-stationary enhancement method (4$^{\text{th}}$ and 5$^{\text{th}}$ columns), projection without enhancement (2$^{\text{th}}$ column) and Ma *et al.* [22] (3$^{\text{rd}}$ column) for four original images (1$^{\text{st}}$ column). The white color in the static text mask means detected text-like regions while green color means detected moving regions. The proposed method sharpen text-like regions with a stronger strength than that of sharpening moving regions.

Table 7.5: Motion artifacts measurements comparison for the proposed method when tested on the *Spinning* video.

Method	$\sigma_{spa} \times 10^{-3}$	$\sigma_{tmp} \times 10^{-3}$
Proposed Method (f_M:28, f_B:32)	**10.18**	**3.75**
Proposed Method (f_M:32, f_B:34)	<u>10.69</u>	<u>3.78</u>
Ma *et al.* ($f_c = 32$) [22]	<u>10.69</u>	<u>3.78</u>
Ma *et al.* ($f_c = 34$) [22]	11.53	3.99

7.4 Content-Adaptive Non-stationary Projector Resolution Enhancement Results

In this section, the content-adaptive non-stationary filtering method introduced in Chapter 6 is tested on the VPAD dataset. Both the quantitative results and qualitative results show the effectiveness of the proposed method.

Figure 7.11 shows a comparison of the proposed content-adaptive method with the state-of-the-art [22] on four VPAD test images. Figure 7.12 shows the PSNR and MSE of the content-adaptive non-stationary enhancement method and Ma *et al.* [22] method over the VPAD video clips, respectively. Over all the test videos, the PSNR and MSE values of the proposed method is always better than that of Ma *et al.* [22]. This means that the proposed method can enhance the projector resolution more while bringing in less additional noise or bias.

7.5 Summary

In this Chapter, the text enhancement method, motion enhancement methods, and comprehensive content-adaptive resolution enhancement method are evaluated qualitatively and quantitatively on our VPAD dataset. As a result, the proposed schemes all offer improved visual quality over projection without enhancement as well as compared to a recent state-of-the-art enhancement method.

[3]Christie Matrix StIM WQ simulation projector

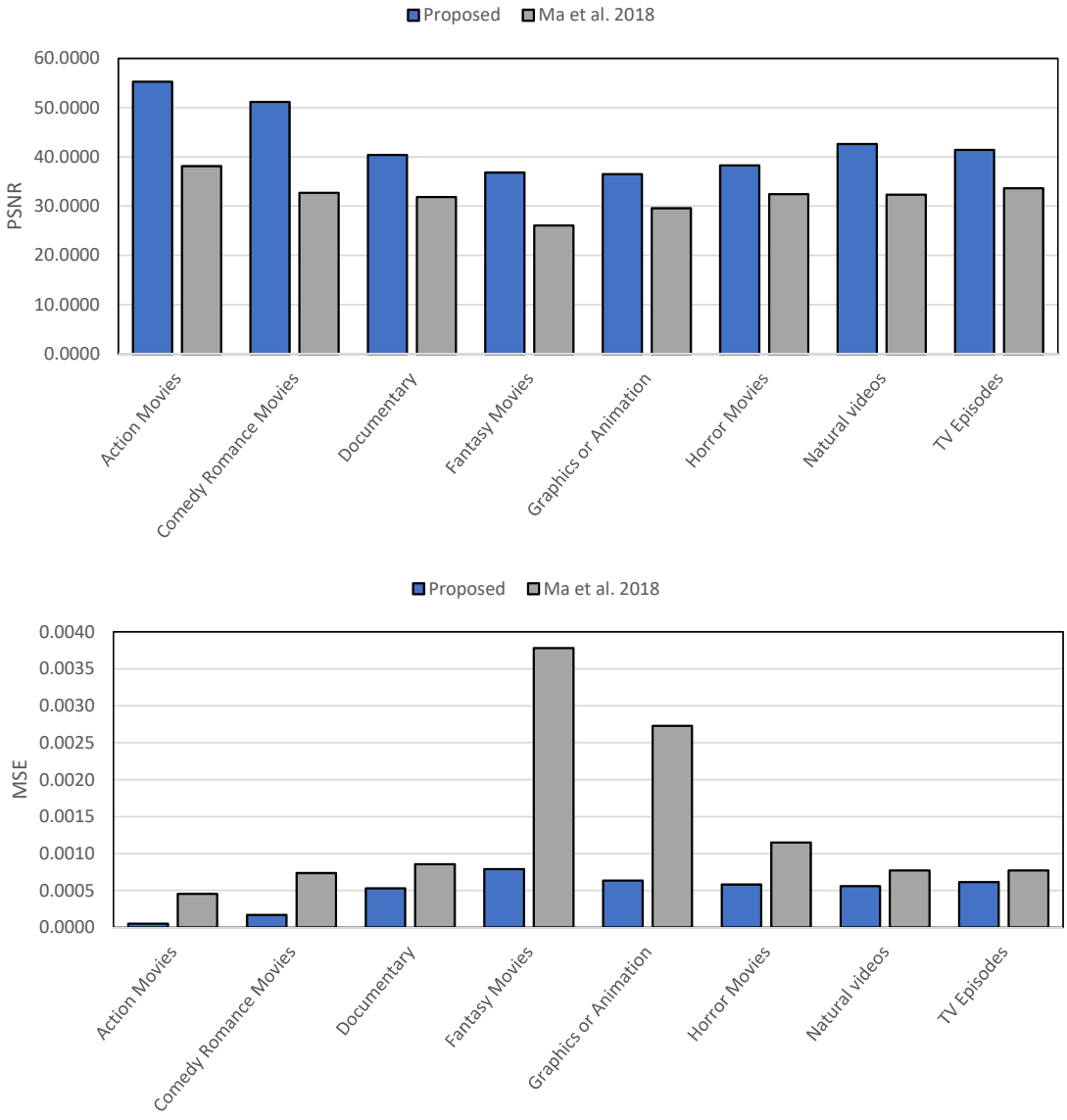


Figure 7.12: Qualitative results of content-adaptive non-stationary enhancement method and Ma *et al.* [22]. the PSNR and MSE values of the proposed method is always better than that of Ma *et al.* [22].

Chapter 8

Conclusion

This chapter summarizes the contributions of this thesis, discusses the impact of this work, and lists potential areas for future research.

8.1 Summary of Thesis and Contributions

In this thesis, a robust projector-based text-like region enhancement scheme, a robust projector-based motion enhancement scheme, and a robust projector-based content-adaptive enhancement scheme are introduced in Chapter 4, 5 and 6. In the text enhancement scheme proposed in Chapter 4, three effective and efficient text-like region detection methods are proposed to classify every pixel into text-like or background class based on the local statistics of high dynamic range regions [1] and the bi-modal characteristic of text-like regions. Two class-dependent Wiener deconvolution kernels of different cutoff frequencies are used in order to sharpen the text-like regions higher than the background ones. Experimental results are conducted on four challenging images and shown that the proposed scheme offers better visual quality than that obtained by projection without enhancement and a recent state-of-the-art enhancement method.

Then, a robust projector-based moving-content enhancement scheme is introduced in Chapter 5. In this scheme, the optical flow-based motion estimation method is proposed [2], and the motion enhancement kernels are generated using directional motion vectors, and then applied to regions with flickering artifacts. It is demonstrated that the proposed motion estimation approach produced robust dense motion vectors and our final results weakened the artifacts appearing by movement in the video. In order to reduce the computational cost and make the computation run more efficiently in projector hardware, a

hypothesis-testing based motion enhancement method is proposed. The moving regions are obtained by computing local statistics to classify every pixel into moving or background class. Two class-dependent Wiener deconvolution filters were used in order to differently enhance motion and background, to avoid the temporal aliasing problem. Experimental results conducted on four videos have shown that the proposed scheme offers better visual quality than that obtained by projection without enhancement and a recent state-of-the-art enhancement method, and efficient in hardware.

The final proposed content-adaptive enhancement scheme is proposed in Chapter 6 by using a novel non-stationary scheme in which the element-wise multiplication between the filtered frames and their corresponding smoothed masks is employed, and then the normalized weighted average is used to obtain the enhanced frame that is ready for projection. This content-adaptive enhancement offers sharpening the high-contrast regions higher than the background ones, while avoids over-sharpening moving regions which may cause temporal motion artifacts. As a result, the proposed scheme offers better visual quality for projected video frames than that of projection without enhancement.

8.2 Impact and Future Work

There are many works in multiple projector projection [80, 81]. It will be interesting if we can apply the content-adaptive resolution enhancement scheme to multi-projector projection.

Besides, the proposed method uses Wiener deconvolution filter introduced in Section 2.2 to enhance the image. However, Wiener deconvolution filter relies on an accurate PSF [25] to inverse the blurring due to the projector-lens system. In this thesis, we assume the PSF is static and applies to all pixels while in practice, the true PSF is not that simple. Hence, in the future, finding a different way to measure the PSF is worthy doing.

Moreover, there are no existing evaluation metrics [82] effective for artifacts measurement. Thus, in the future, it is also valuable to find a good assessment metrics for superimposition-based resolution-enhanced images.

References

[1] X. Hu, M. A. Naiel, Z. Azimifar, I. Ben Daya, M. Lamm, and P. Fieguth. Text enhancement in projected imagery. *Journal of Computational Vision and Imaging Systems*, 4(1):3, Dec. 2018.

[2] X. Hu, A. Ma, A. Gawish, M. Lamm, and P. Fieguth. Motion detection in high resolution enhancement. *Journal of Computational Vision and Imaging Systems*, 3, 2017.

[3] M. Ashdown, P. Tuddenham, P. Robinson, and C. Mller-Tomfelde. High-resolution interactive displays. pages 71–100, 2010.

[4] O. Schreer, I. Feldmann, C. Weissig, P. Kauff, and R. Schafer. Ultrahigh-resolution panoramic imaging for format-agnostic video production. *Proceedings of the IEEE*, 101:99–114, 2013.

[5] R. Raskar, M. S. Brown, Ruigang Yang, Wei-Chao Chen, G. Welch, H. Towles, B. Scales, and H. Fuchs. Multi-projector displays using camera-based registration. In *Proceedings Visualization '99 (Cat. No.99CB37067)*, pages 161–522, 1999.

[6] W. Allen and R. Ulichney. 47.4: Invited paper: Wobulation: Doubling the addressed resolution of projection displays. *SID Symposium Digest of Technical Papers*, 36:1514–1517.

[7] M. Brown, A. Majumder, and R. Yang. Camera-based calibration techniques for seamless multiprojector displays. *IEEE Transactions on Visualization and Computer Graphics*, 11:193–206, 2005.

[8] N. Damera-Venkata and N. L. Chang. Realizing super-resolution with superimposed projection. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[9] N. Damera-Venkata and N. L. Chang. Display supersampling. *ACM Trans. Graph.*, 28:9:1–9:19, 2009.

[10] E. Barshan, M. Lamm, C. Scharfenberger, and P. Fieguth. Resolution enhancement based on shifted superposition. *SID Symposium Digest of Technical Papers*, 46:514–517.

[11] A. P. Dhawan, R. M. Rangayyan, and R. Gordon. Image restoration by wiener deconvolution in limited-view computed tomography. *Appl. Opt.*, 24:4013–4020, 1985.

[12] A. Wong, Y. Li, M. Lamm, and H. Sekkati. Device, system and method for content-adaptive resolution-enhancement, 2016.

[13] S. C. Park, M. K. Park, and M. G. Kang. Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, 20:21–36, 2003.

[14] 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*, volume 33, pages 1254–1257, 2002.

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

[16] N. Damera-Venkata and N. L. Chang. On the resolution limits of superimposed projection. In *2007 IEEE International Conference on Image Processing*, volume 5, pages V – 373–V – 376, Sep. 2007.

[17] B. Sajadi, D. Qoc-Lai, A. H. Ihler, M. Gopi, and A. Majumder. Image enhancement in projectors via optical pixel shift and overlay. In *IEEE International Conference on Computational Photography (ICCP)*, pages 1–10, 2013.

[18] F. Berthouzoz and R. Fattal. Resolution enhancement by vibrating displays. *ACM Trans. Graph.*, 31:15:1–15:14, 2012.

[19] P. Didyk, E. Eisemann, T. Ritschel, K. Myszkowski, and H. Seidel. Apparent display resolution enhancement for moving images. *ACM Transactions on Graphics (Proceedings SIGGRAPH 2010, Los Angeles)*, 29, 2010.

[20] N. Damera-Venkata and N. L. Chang. Display supersampling. *ACM Trans. Graph.*, 28:9:1–9:19, 2009.

[21] D. G. Aliaga, Y. H. Yeung, A. Law, B. Sajadi, and A. Majumder. Fast high-resolution appearance editing using superimposed projections. *ACM Trans. Graph.*, 31:13:1–13:13, April 2012.

[22] 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*, 49:831–834, 2018.

[23] M. Hasan and M. R. El-Sakka. Structural similarity optimized wiener filter: A way to fight image noise. In Mohamed Kamel and Aurélio Campilho, editors, *Image Analysis and Recognition*, pages 60–68, 2015.

[24] A. Klug and R. A. Crowther. Three-dimensional image reconstruction from the viewpoint of information theory. *Nature Structural and Molecular Biology*, 238, 1972.

[25] M. S. Brown, P. S., and T. Cham. Image pre-conditioning for out-of-focus projector blur. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 1956–1963, 2006.

[26] Q. Ye and D. Doermann. Text detection and recognition in imagery: A survey. *IEEE Trans. on Pattern Anal. Mach. Intell.*, 37:1480–1500, 2015.

[27] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. J. Wu, and A. Y. Ng. Text detection and character recognition in scene images with unsupervised feature learning. In *2011 International Conference on Document Analysis and Recognition*, pages 440–445, 2011.

[28] S. Shetty, A. S. Devadiga, S. S. Chakkaravarthy, and K. A. V. Kumar. Ote-ocr based text recognition and extraction from video frames. In *2014 IEEE 8th International Conference on Intelligent Systems and Control (ISCO)*, pages 229–232, 2014.

[29] J. Gllavata, R. Ewerth, and B. Freisleben. Text detection in images based on unsupervised classification of high-frequency wavelet coefficients. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 1, pages 425–428, 2004.

[30] B. Freisleben, R. Ewerth, and J. Gllavata. A text detection, localization and segmentation system for ocr in images. In *Multimedia Software Engineering, International Symposium on(ISMSE)*, volume 00, pages 310–317, 2004.

[31] D. Chen, J. Odobez, and H. Bourlard. Text detection and recognition in images and video frames. *Pattern Recognition*, 37:595–608, 2004.

[32] M. R. Lyu, J. Song, and M. Cai. A comprehensive method for multilingual video text detection, localization, and extraction. *IEEE Transactions on Circuits and Systems for Video Technology*, 15:243–255, 2005.

[33] L. Wenyin, J. Xi, X. Chen, X. Hua, and H. Zhang. A video text detection and recognition system. In *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.(ICME)*, volume 00, page 222, 2001.

[34] L. Agnihotri and N. Dimitrova. Text detection for video analysis. In *Proceedings IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL'99)*, pages 109–113, 1999.

[35] S. Xu, J. McCusker, M. Schultz, and M. Krauthammer. Improving ocr performance in biomedical literature retrieval through preprocessing and postprocessing. 2008.

[36] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22:761–767, 2004.

[37] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod. Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In *Proc. 18th IEEE Int. Conf. on Image Processing*, pages 2609–2612, 2011.

[38] M. Donoser and H. Bischof. Efficient maximally stable extremal region (MSER) tracking. In *Proc. IEEE Conf. on Computer Vision and Pattern Recogn.*, volume 1, pages 553–560, 2006.

[39] K. Huang X-C Yin, X-W Yin and H. Hao. Robust text detection in natural scene images. *IEEE Trans. on Pattern Anal. Mach. Intell.*, 36:970–983, 2014.

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

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

[42] M. Smith and T. Kanade. Video skimming for quick browsing based on audio and image characterization. Technical report, Carnegie Mellon University, Pittsburgh, PA, July 1995.

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

[44] 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.*, 15:243–255, 2005.

[45] R. Nakayama and I. Motoyoshi. Sensitivity to acceleration in the human early visual system. *Frontiers in Psychology*, 8:925, 2017.

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

[47] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In Tomás Pajdla and Jiří Matas, editors, *Computer Vision - ECCV 2004*, pages 25–36, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[48] M. Maule, J. L. D. Comba, R. Torchelsen, and R. Bastos. Transparency and anti-aliasing techniques for real-time rendering. In *2012 25th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials*, pages 50–59, 2012.

[49] F. Arif and M. Akbar. A new approach for anti-aliasing raster data in air borne imagery. In *2005 International Conference on Information and Communication Technologies*, pages 90–93, 2005.

[50] J. Yang and H. Li. Dense, accurate optical flow estimation with piecewise parametric model. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1019–1027, 2015.

[51] D. Fortun, P. Bouthemy, and C. Kervrann. Optical flow modeling and computation: A survey. *Computer Vision and Image Understanding*, 134:1–21, 2015. Image Understanding for Real-world Distributed Video Networks.

[52] A. Bruhn, J. Weickert, and C. Schnörr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, 61:211–231, Feb 2005.

[53] C. Liu. Beyond pixels: Exploring new representations and applications for motion analysis. 2009.

[54] R. Kalman. A new approach to linear filtering and prediction problems. 82D:35–45, 1960.

[55] S. Y. Chen. Kalman filter for robot vision: A survey. *IEEE Transactions on Industrial Electronics*, 59:4409–4420, 2012.

[56] T. J. Broida, S. Chandrashekhar, and R. Chellappa. Recursive 3-d motion estimation from a monocular image sequence. *IEEE Transactions on Aerospace and Electronic Systems*, 26:639–656, 1990.

[57] N. P. Papanikolopoulos, P. K. Khosla, and T. Kanade. Visual tracking of a moving target by a camera mounted on a robot: a combination of control and vision. *IEEE Transactions on Robotics and Automation*, 9:14–35, 1993.

[58] S. I. Roumeliotis and G. A. Bekey. Distributed multirobot localization. *IEEE Transactions on Robotics and Automation*, 18:781–795, 2002.

[59] C. Kuo, C. Chao, and C. Hsieh. A new motion estimation algorithm for video coding using adaptive kalman filter. *Real-Time Imaging*, 8:387–398, October 2002.

[60] S. Chung, C. Kuo, and P. Shih. Rate-constrained motion estimation using kalman filter. *Journal of Visual Communication and Image Representation*, 17:929–946, 2006.

[61] A. Singh. Incremental estimation of image-flow using a kalman filter. In *Proceedings of the IEEE Workshop on Visual Motion*, pages 36–43, 1991.

[62] J. R. Cooper and R. O. Hastings. Kalman filtering from a phase based optical flow operator. In Abdul Sattar, editor, *Advanced Topics in Artificial Intelligence*, pages 77–86, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg.

[63] B. Haskell. Frame-to-frame coding of television pictures using two-dimensional fourier transforms (corresp.). *IEEE Transactions on Information Theory*, 20:119–120, 1974.

[64] J. Jain and A. Jain. Displacement measurement and its application in interframe image coding. *IEEE Transactions on Communications*, 29:1799–1808, 1981.

[65] J. Skowronski. Pel recursive motion estimation and compensation in subbands. *Signal Processing: Image Communication*, 14:389–396, 1999.

[66] K. N. OGLE. The perception of the visual world. james j. gibson; leonard carmichael, ed. boston: Houghton mifflin. *Science*, 113:535–535, 1951.

[67] R. Szeliski. Algorithms and applications. 2010.

[68] B. K.P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[69] P. Fieguth. Statistical image processing and multidimensional modeling. *Real-Time Imaging*.

[70] D. Stricker G. Bleser. Extended kalman filter. *lecture notes, Systems Design SS 2014 - Computer Vision: Object and People Tracking, University of Kaiserslautern, delivered 2 September 2010.*

[71] S. D. Levy. The extended kalman filter: An interactive tutorial. *Lee University, delivered 19 December 2016.*

[72] D. Lelescu and D. Schonfeld. Statistical sequential analysis for real-time video scene change detection on compressed multimedia bitstream. *IEEE Transactions on Multimedia*, 5(1):106–117, 2003.

[73] A. Walha, A. Wali, and A. M. Alimi. Video stabilization with moving object detecting and tracking for aerial video surveillance. *Multimedia Tools and Applications*, 74:6745–6767, 2015.

[74] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.

[75] J. J. Quinlan and C. J. Sreenan. Multi-profile ultra high definition (UHD) AVC and HEVC 4K dash datasets. In *Proc. ACM-MMSYS*, 2018.

[76] https://archive.org. Last retrieved Mar. 15th, 2019.

[77] J. P. Lewis S. Roth M. J. Black S. Baker, D. Scharstein and R. Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92:1–31, Mar 2011.

[78] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2432–2439, 2010.

[79] D. Sun, S. Roth, and M. J. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision*, 106:115–137, Jan 2014.

[80] I. Kauvar, S. J. Yang, L. Shi, I. McDowall, and G. Wetzstein. Adaptive color display via perceptually-driven factored spectral projection. *ACM Trans. Graph.*, 34:165:1–165:10, 2015.

[81] A. Hryniowski, I. Ben Daya, A. Gawish, M. Lamm, A. Wong, and P. Fieguth. Multi-projector resolution enhancement through biased interpolation. pages 190–197, 2018.

[82] S. A. J. Hansen, M. N. Akram, J. Y. Hardeberg, and M. Pedersen. Preferred image quality metric for shifted superimposition-based resolution-enhanced images. *Journal of Electronic Imaging*, 27:1–13–13, 2018.

[83] R. Zabih, J. Miller, and K. Mai. A feature-based algorithm for detecting and classifying scene breaks. In *Proceedings of the Third ACM International Conference on Multimedia*, pages 189–200, 1995.

[84] A. Gonzlez, L. M. Bergasa, J. J. Yebes, and S. Bronte. Text location in complex images. In *Proc. 21st Int. Conf. on Pattern Recogn.*, pages 617–620, 2012.

[85] Y. Li and H. Lu. Scene text detection via stroke width. In *Proc. 21st Int. Conf. on Pattern Recogn.*, pages 681–684, 2012.

[86] L. Neumann and J. Matas. Real-time scene text localization and recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recogn.*, pages 3538–3545, 2012.

[87] X. Wang, Y. Song, and Y. Zhang. Natural scene text detection with multi-channel connected component segmentation. In *2013 12th Int. Conf. on Document Analysis and Recogn.*, pages 1375–1379, 2013.

[88] J. Gllavata, R. Ewerth, and B. Freisleben. A text detection, localization and segmentation system for ocr in images. In *IEEE Sixth Int. Symp. on Multimedia Software Engineering*, pages 310–317, 2004.

[89] C. Yi and Y. Tian. Text string detection from natural scenes by structure-based partition and grouping. *IEEE Trans. on Image Processing*, 20:2594–2605, 2011.

[90] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang. Framework for performance evaluation of

face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Trans. on Pattern Anal. Mach. Intell.*, 31:319–336, 2009.

[91] J. Gllavata, R. Ewerth, and B. Freisleben. Finding text in images via local thresholding. In *Proc. 3rd IEEE Int. Symp. on Signal Processing and Information Technology*, pages 539–542, 2003.

[92] F. Arif and M. Akbar. A new approach for anti-aliasing raster data in air borne imagery. In *2005 International Conference on Information and Communication Technologies*, pages 90–93, 2005.

[93] D. Fortun, P. Bouthemy, and C. Kervrann. Optical flow modeling and computation: A survey. *Computer Vision and Image Understanding*, 134:1–21, 2015. Image Understanding for Real-world Distributed Video Networks.

[94] L.N. Thibos, D.J. Walsh, and F.E. Cheney. Vision beyond the resolution limit: Aliasing in the periphery. *Vision Research*, 27:2193 – 2197, 1987.

[95] R. Zabih, J. Miller, and K. Mai. A feature-based algorithm for detecting and classifying scene breaks. In *Proceedings of the Third ACM International Conference on Multimedia*, MULTIMEDIA '95, pages 189–200, New York, NY, USA, 1995. ACM.

[96] J. C. Alvarez. Estimation of the longitudinal and lateral velocities of a vehicle using extended kalman filters. 2018.

[97] S. J. Lee, Y. Motai, and M. Murphy. Respiratory motion estimation with hybrid implementation of extended kalman filter. *IEEE Transactions on Industrial Electronics*, 59:4421–4432, 2012.

[98] K. Zindler, N. Geiss, K. Doll, and S. Heinlein. Real-time ego-motion estimation using lidar and a vehicle model based extended kalman filter. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 431–438, 2014.

[99] M. Bajaj and B. Lall. Enhanced motion estimation using kalman filter. *IETE Journal of Research*, 58:171–175, 2012.

[100] X. Hu, M. A. Naiel, Z. Azimifar, M. Lamm, and P. Fieguth. Robust visual enhancement of moving contents in projected imagery. *SID Symposium Digest of Technical Papers*, 2019.