Model-based Regularization for Video Super-Resolution

MODEL-BASED REGULARIZATION FOR VIDEO SUPER-RESOLUTION

BY

HUAZHONG WANG, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

Master of Applied Science (2009) McMaster University
(Electrical & Computer Engineering) Hamilton, Ontario, Canada

TITLE:	Model-based Regularization for Video Super-Resolution
AUTHOR:	Huazhong Wang B.Sc., (Electronic Engineering and Information Science) University of Science and Technology of China, Hefei, China
SUPERVISOR:	Dr. Xiaolin Wu
NUMBER OF PAGES:	xi, 75

To my beloved family

Abstract

In this thesis, we reexamine the classical problem of video super-resolution, with an aim to reproduce fine edge/texture details of acquired digital videos. In general, the video super-resolution reconstruction is an ill-posed inverse problem, because of an insufficient number of observations from registered low-resolution video frames. To stabilize the problem and make its solution more accurate, we develop two video super-resolution techniques: 1) a 2D autoregressive modeling and interpolation technique for video super-resolution reconstruction, with model parameters estimated from multiple registered low-resolution frames; 2) the use of image model as a regularization term to improve the performance of the traditional video super-resolution algorithm.

We further investigate the interactions of various unknown variables involved in video super-resolution reconstruction, including motion parameters, high-resolution pixel intensities and the parameters of the image model used for regularization. We succeed in developing a joint estimation technique that infers these unknowns simultaneously to achieve statistical consistency among them.

Acknowledgements

I feel an immense gratitude to my supervisor Dr. Xiaolin Wu for his continuous academic guidance and great patience throughout the work for my master's degree. I appreciate the opportunity to gain experience by working with an individual that constantly strives for quality, professionalism and perfection. The knowledge acquired from Dr. Wu is indispensable and will be tremendously beneficial in my future career.

I would also like to thank my examiners, Dr. Sorina Dumitrescu and Dr. Jian-Kang Zhang, for their time reviewing my thesis and providing valuable feedback. I appreciate the friendly assistance and expert technical support provided by Cheryl and Cosmin, respectively.

Furthermore, I wish to thank all my colleagues and friends with whom I spent two years at McMaster. Their friendship, encouragement and support allowed me to enjoy a challenging but wonderful experience in Canada.

Last but not least, I would like to thank my family. Their great love and support is more than words can explain, and it truly provided me with the momentum needed to persevere through my struggles in the world.

Notation and abbreviations

2D	Two-dimension
AWGN	Additive White Gaussian Noise
AR	Autoregressive
BMA	Block-Matching Algorithm
CCD	Charge-Coupled Device
CMOS	Complementary Metal Oxide Semiconductor
DCT	Discrete Cosine Transform
DPI	Dots Per Inch
HD	High-definition
HMRF	Huber-Markov Random Field
HR	High-resolution
LR	Low-resolution
LS	Least Squares
LSI	Linear Space-Invariant

MAP	Maximum *a posteriori*
ML	Maximum Likelihood
MMSE	Minimum Mean Squared Error
PAR	Piecewise Autoregressive
PDE	Partial Differential Equation
POCS	Projection Onto Convex Sets
PPI	Pixels Per Inch
PSF	Point Spread Function
PSNR	Peak Signal-to-Noise Ratio
SAD	Sum of Absolute Differences
SNR	Signal-to-Noise Ratio
SSD	Sum of Squared Differences
TV	Total Variation
VSR	Video Super-Resolution

Contents

List of Figures

Chapter 1

Introduction and Problem Statement

1.1 The Problem and Motivation

In our technology era digital images and videos provide important and ubiquitous means of visual communication. As what is said, 'A picture is worth a thousand words', digital images, in both still and moving forms, convey knowledge and information in a more intuitive and convenient manner than texts. The growth of digital visual media has been phenomenal with the rapid development and deployment of digital imaging and communication technologies. According to an official announcement from Google in February 2005, the number of still images indexed by Google Images search engine amounts to be 1.1 billion [12]. Three years later, Facebook announces that they host 10 billion images till October 2008 [8]. In addition to the steadily growing quantity above, the quest for higher quality of digital images in a variety of applications has never abated.

One of the most important attributes of a digital image/video is its resolution. The resolution is often used as a measure of the size and visual quality (e.g., clarity) of digital image/video. In general, given a scene, the higher the image resolution, the more and finer details an image/video contains. In the past decade, the resolution of digital imaging devices has steadily improved thanks to advances in semiconductor and sensor technologies. At present even inexpensive mainstream consumer cameras can have eight or more millions of pixels.

However, for many high-end and professional applications, such as those in medicine, biology, astronomy, military, visual arts, etc., the resolution of digital image/video will never be high enough. Human pursue of knowledge is endless and we always want to push the envelop and image ever minuscule structures and details in nature. Since many image signals are band unlimited, by Nyquist sampling theorem the sampling frequency (i.e., resolution) of digital image/video has to be sufficiently high to completely recover the underlying continuous light field. Unfortunately, the image/video resolution is bounded by some hard physical limits. First, diffraction limit of optical lens system prevents the infinite resolving of continuous image signals [4]. Second, the inherently finite nature of digital sensor technologies and the imperfection of manufacturing process place a cap on the achievable image resolution. Most digital images are acquired by an array of semiconductor sensors such as Charge-Coupled Device (CCD) and Complementary Metal Oxide Semiconductor (CMOS). These types of devices are fast approaching the density limit in microelectronics. As the image resolution gets higher and higher, namely, pixels smaller and smaller, the amount of light intercepted by each pixel diminishes, reducing the signal strength. To make the matter worse, more densely packed sensors induce a greater amount of electronic inferences between the neighboring pixel sensors. Consequently, the signal-to-noise ratio (SNR) of the acquired image decreases in resolution.

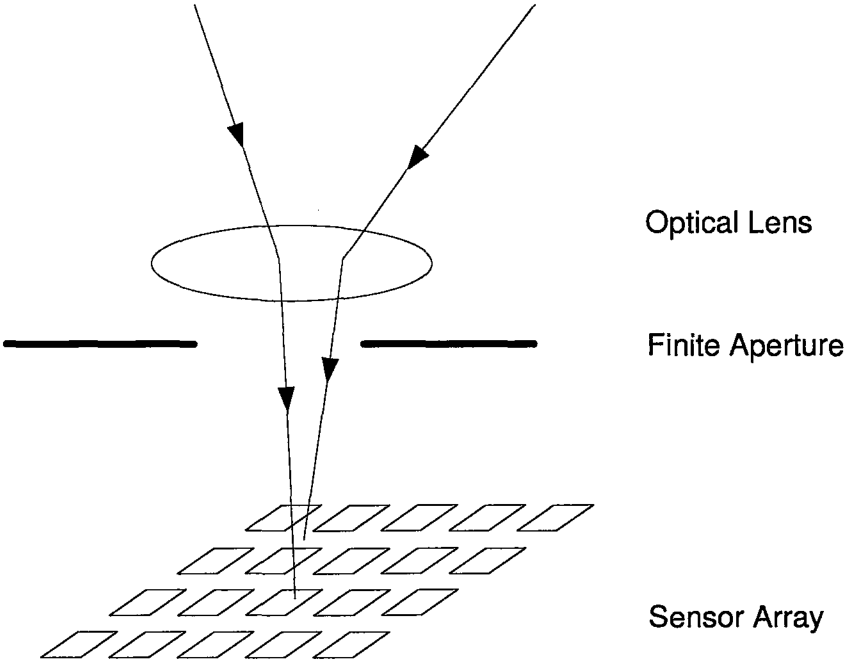


Figure 1.1: Illustration of the digital optical imaging system

Due to the aforementioned limits of the digital imaging technologies and systems, it is unlikely that newer imaging devices in the future, by themselves, can completely meet the resolution and precision requirements of many scientific, medical and military applications at present and in the future. In this case, the only alternative is to compensate for the inadequacy of sensor resolution via digital image/video processing after the data acquisition. Image interpolation and video super-resolution (VSR) techniques are commonly used to improve the native sensor resolution of imaging devices. They aim to recover a high-resolution image or video frame from the observed lower-resolution version.

VSR techniques are also useful to improve cost effectiveness of video products and services. In a large-scale video surveillance system, for example, only low-resolution video cameras can be economically deployed because of the sheer number of cameras involved. Users can rely on VSR techniques to enhance the video quality and achieve a similar system performance as more expensive high-resolution cameras can provide.

Another important application of VSR techniques is to upconvert existing low-resolution low-quality video contents to higher resolution and higher quality. The needs for video resolution upconversion are increasing and become ever pressing. Nowadays high-definition television sets, computer monitors, and blu-ray players are commonplace. But many old valuable digital contents are of standard definition, such as standard VCD and DVD formats. This quality gap between the input materials and output devices can only be bridged by VSR techniques. The market potential is huge for VSR products that can rebuild the large connection of old movies and television programs for modern high quality output devices.

1.2 An Introduction to Video Super-Resolution

As mentioned above, one of the most important quality metrics of digital images is the resolution. In image/video processing, the terminology of resolution can refer to two different notions namely, pixel resolution and spatial resolution. Pixel resolution, in brief, refers to the number of pixels in a digital image. For example, an image that is 1000 pixels in width and 800 pixels in height (denoted by 1000×800) has a total of 0.8 million pixels. However, high pixel resolution does not necessarily correspond to high visual quality. Instead, spatial resolution is a measure of image fidelity and quality, in particular for high-frequency features, such as edges and textures. Spatial resolution refers to the pixel density of a digital image. It is defined to be the number of sampling points per unit length/area of continuous image signals. In this sense, spatial resolution is commonly measured in Dots Per Inch (DPI), Pixels Per Inch (PPI) or Per Square Inch. Manufacturers of devices such as digital scanners, printers and monitors, often take spatial resolution as a measure of the device capability to resolve details of optical signals. For example, nowadays, typical office scanners can have a resolution of 1200 dpi or higher. For such image acquisition devices, the higher the spatial resolution, the finer edge/texture details can be resolved.

From the perspective of signal processing, the problem of upconverting the resolution of an acquired digital image can be interpreted as re-sampling of the original continuous two-dimensional (2D) image signal at a higher spatial sampling frequency. It is equivalent to the problem of reconstructing the continuous image signal from a set of observed (measured) discrete samples (pixels). According to the Nyquist-Shannon sampling theorem, those signal components that have frequency lower than the Nyquist frequency can be exactly reproduced. It indicates that, low-frequency image signal which in general manifests as smooth area or large-scale edge/texture, can be reconstructed easily. In other words, the challenge of image resolution upconversion lies in reproducing the high-frequency components of the image signal that exceed the Nyquist limit. Reproduction of these high-frequency components namely, edge details and fine textures, provides the possibility to improve the visual quality of acquired images. As a matter of fact, due to the point spread function (PSF) of photoelectric sensors that plays a role of low-pass filtering, many of the high-frequency components are attenuated to avoid aliasing[1]. In practice, any attempts to reproduce these high-frequency components are prone to artifacts. The most common artifacts due to reconstruction error, are visually noticeable blur, jaggies and aliasing. Human vision system is highly sensitive to such artifacts that arise in the areas of edges/textures. These arfacts degrade the fidelity and quality of images, and hence should be avoided in image resolution upconversion.

In the past decade, intensive research efforts have been devoted to developing resolution upconversion techniques that can reproduce the high-frequency image signals free of artifacts. In terms of image form, all published techniques to date can be categorized into two classes namely, single-image based methods and multi-image based methods. The former refers to those methods that upconvert image resolution from a single still image of lower resolution. As most of them improve the image resolution by interpolating, they are also known as still image interpolation techniques. Among these techniques in the literature are Bicubic interpolation[18], directional interpolation[43] and soft-decision adaptive interpolation[44]. Due to high efficiency and low computational cost, image interpolation techniques are widely used to resize digital images and video frames. The multi-image based methods are commonly referred as the aforementioned VSR techniques. In general, these methods involve the state-of-the-art image registration, and are used to deal with video frames. They take advantage of pixel information from multiple video frames, and hence can achieve superior performance against still image interpolation methods. However, compared with still image interpolation counterparts, VSR techniques are much more complicated and often undergo high computational cost. In this thesis, our study is focused on developing VSR methods for the purpose of image resolution upconversion.

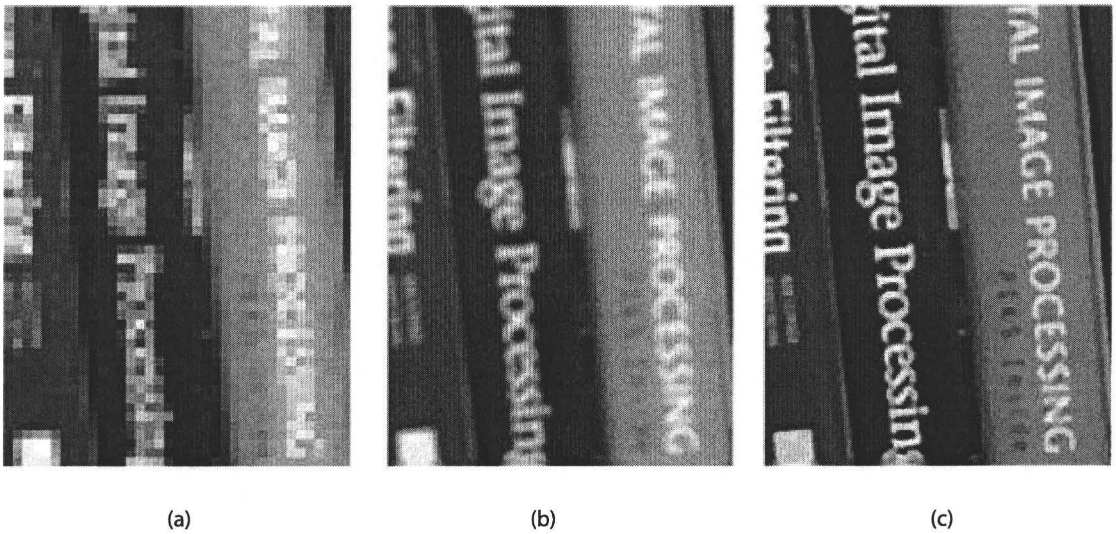


Figure 1.2: An example of video super-resolution results from [9]. (a) one of observed low-resolution frames; (b) a high-resolution frame reconstructed from multiple observed low-resolution frames; (c) a deblurred high-resolution frame from (b).

As discussed above, VSR is a technology of reconstructing one or more high-resolution (HR) video frames from a sequence of degraded low-resolution (LR) video frames. It has a widespread application in areas such as medical imaging (i.e., CAT, MRI, etc.), satellite imaging, enlarging consumer photographs, video surveillance, etc. Fig. 1.2(c) shows an example of VSR produced high-resolution video frame, in contrast with one of the observed low-resolution frames as shown in (a). The frame shown in (b) is an intermediate result in which the PSF effect of photoelectric sensors has not been eliminated. In the literature, the process of eliminating the PSF effect is also known as deblurring or deconvolution.

The basic idea behind the VSR technique is the fusing of multiple low-resolution frames with subpixel shifts to reconstruct high-resolution video frame(s)[35]. In the scenario of a digital video camera system, a video sequence is produced by sampling the continuous image signal of a scene at a constant frame rate, e.g., 24 or 30 frames per second (fps). The acquired adjacent video frames are naturally shifted at subpixel displacements, given that the scene does not change and objects in the scene move with subpixel increments with respect to the video camera. Consequently, by registering these adjacent frames, VSR techniques can substantially increase the pixel density i.e., spatial resolution of these acquired video frames. Fig. 1.3 illustrates the fusion of multiple frames at subpixel displacements to produce an HR frame. The observed low-resolution frames are first registered against a reference frame. This procedure employs image registration to form a high-resolution grid, and therefore it is also known as motion-based image interpolation. Due to arbitrary motions of the moving objects, registered pixels on the high-resolution grid may not be distributed at regular pixel sites. Then, they are mapped onto another high-resolution grid where pixels are distributed at regular pixel sites. This mapping is also known as grid-based image interpolation, for these resulting pixels are interpolated from the prior registered pixels. In addition to the direct (i.e., one-pass) mapping, this step can also be an iterative procedure that projects the interpolated pixels back to the registered grid and checks the validity of the grid-based interpolation iteratively. At each iteration, the interpolated pixels on the high-resolution grid are updated. The iterative procedure converges when meeting a required threshold for the back projection error. The next operation on the reconstructed high-resolution frame is deblurring that eliminates the PSF effect of photoelectric sensors. It should be pointed out that, in the case of a scene moving with integer pixel units, VSR techniques can not help improve the spatial resolution of video frames, for the adjacent frames contain the same pixel information.

In addition to upconverting image resolution of an acquired video sequence, another capability of the VSR technique is to alleviate image noise. The most common image noise originates in image acquisition devices. It can significantly degrade the image quality. Moreover, this type of image noise is independent of image signals, and can be modeled as additive white Gaussian noise (AWGN). By fusing together multiple adjacent frames, the VSR technique is capable of effectively suppressing the additive image noise in video frames.

1.3 Review of Video Super-Resolution

In the early 1980s, Tsai and Huang proposed a frequency domain approach in employing multiple noiseless down-sampled frame to enhance the resolution of a frame[40]. This method, distinguished from conventional counterparts that use a still image for

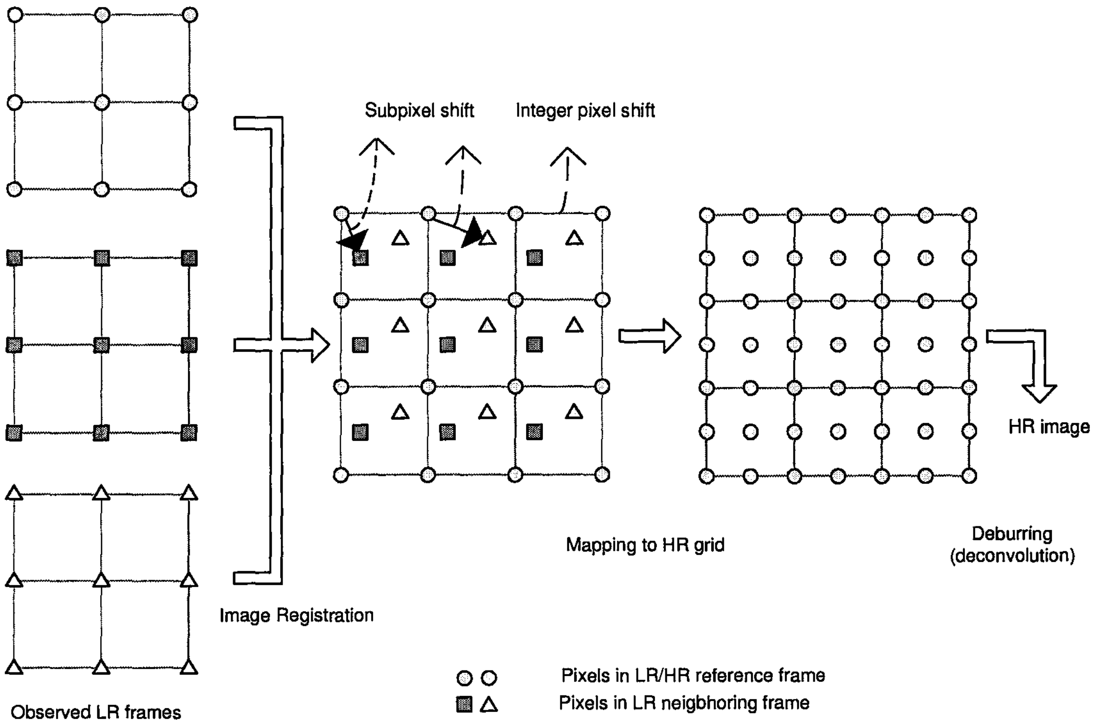


Figure 1.3: Illustration of reconstructing a high-resolution video frame from multiple observed low-resolution video frames.

resolution upconversion, is the pioneering work on VSR reconstruction. Since then, VSR as a research topic has been receiving so much attention. Currently, the VSR problem is generalized as reconstructing an image (frame) of higher resolution from several down-sampled and degraded images (frames). Up to now, all published VSR algorithms can be categorized into two classes namely, frequency domain methods and spatial domain methods[3].

After Tsai and Huang[40], Kim *et al.* proposed a recursive frequency domain algorithm to reproduce HR frames from a noisy and down-sampled image sequence[20, 21, 19]. The authors first took into account noise and spatial blurring and utilized the Tikhonov regularization for the image reconstruction problem. Meanwhile, Srinivas and Srinath proposed an algorithm based on a minimum mean squared error (MMSE) approach for the multi-image restoration problem[38]. Later, Rhee and Kang developed a discrete cosine transform (DCT) based frequency method[31]. The above algorithms are capable of dealing with linear space-invariant (LSI) blur, as well as homogeneous additive noise. Moreover, they can achieve high computational efficiency. Nevertheless, pixel displacements within frames are restricted to global uniform translation motion. Furthermore, they are not able to exploit *a priori* knowledge of spatial domain pixel structures and make the reconstruction of high-resolution image (frame) adapt to local image waveform.

The first spatial domain algorithm based on the iterative back projection (IBP) was proposed by Irani and Peleg for the VSR reconstruction problem[16, 17, 28]. They initialize the iterative process with a guess of the missing high-resolution image. At each iteration, the algorithm projects the temporary results to the observed LR images. By computing the projection error, the algorithm updates the guessed image iteratively and makes it approach the optimal result. The highlight of this algorithm lies in its capacity in dealing with affine geometric warps. Tekalp and Sezan proposed the concept of convex sets that can be used efficiently as a constraint for the ill-posed VSR problem[27, 39]. This approach makes it possible to combine the nonlinear constraint to VSR reconstruction and to apply the projection onto convex sets (POCS) method in VSR. However, POCS method has the disadvantage of high computational cost and slow convergence. Later, another algorithm for VSR reconstruction problem was proposed by Cheeseman *et al.*, by using maximum *a posteriori* (MAP) based on a Gaussian smoothness prior[6]. Schultz and Stevenson suggested an approach by using a MAP estimator with the Huber-Markov Random Field (HMRF) prior[34, 35]. This approach works based on the assumption of averaging box point spread function (PSF) and the additive noise which is assumed to be independent and identical distributed (i.i.d.) Gaussian random variable. All these VSR algorithms have their advantages in different aspects. In 1997, Elad and Feuer proposed a new unified framework for VSR reconstruction by reconstructing from multiple blurred, noisy, and down-sampled observed images. The authors proposed to formulate the VSR reconstruction problem by using sparse matrices from the perspective of maximum likelihood (ML), MAP, and POCS.

The VSR algorithms can also be categorized in some other criteria, other than the domain. For example, there are deterministic and non-deterministic (stochastic) variants of VSR methods. The deterministic methods employ some *a priori* knowledge of the observed image, such as smoothness, to regularize the solution space of the ill-posed VSR problem. Some other VSR algorithms that use the MAP methods treat the VSR reconstruction as a probability estimation problem[13, 36]. They belong to the non-deterministic class. Moreover, VSR methods can also be classified into iterative and non-iterative categories.

1.4 Contributions

My research work, as presented in this thesis, targets at developing image resolution upconversion methods to reproduce fine edge/texture details of acquired digital videos. Our design principle is inspired and motivated by the ability of a piecewise autoregressive (PAR) image model in modeling digital image signals and preserving spatial structure of pixels in still image interpolation. In the proceeding chapters, we present two new VSR methods that provide effective solutions for the VSR reconstruction problem. Both the two VSR methods take advantage of the PAR model, but perform image resolution upconversion in different ways. Simulation results demonstrate that both achieve competitive performance in terms of perceptual quality. The contributions of this thesis are summarized as follows:

- We extend the PAR model to solve the VSR problem. In still image interpolation, PAR model parameters are estimated from a local window of observed low-resolution still image, but applied to reconstruct the underlying high-resolution image. Considering a potential mismatch of the PAR model, we propose to estimate PAR model parameters from observed data of multiple registered low-resolution video frames. Compared with that in still image interpolation, learning of PAR models by the new scheme is much more accurate, and hence significantly reduces the likelihood of model mismatch between the observed low-resolution frame and the underlying high-resolution frame. Then VSR reconstruction is simplified as a still image interpolation problem, and the underlying high-resolution video frames are reconstructed by using the above estimated PAR model parameters. In this case, image registration is implicitly incorporated into the reconstruction of high-resolution video frames. It makes the solution for this problem more robust to motion estimation errors. The new method gains superior performance against its counterpart in terms of both visual quality and peak signal-to-noise ratio(PSNR) measurement.

- We propose to incorporate the PAR model into a regularization term for the inverse VSR problem. The VSR reconstruction problem is formulated via a well-known multi-frame observation model which combines explicit motion estimation. Due to the lack of constraints, the VSR problem is ill posed. Historically, total variation (TV) methods are most commonly used to impose constraints on the solution space of the ill-posed VSR problem. However, TV methods ignore the second and higher order derivatives of image signals, and further can not adapt to local image waveforms. Consequently, VSR methods that employ TV for regularization can not preserve image details but force the smoothness of image signals. By contrast, the PAR model based regularization method can, by spatially varing its parameters, adapt the reconstruction of HR frames to the local image waveforms. As a result, it can effectively regularize solutions for the ill-posed VSR problem and reproduce high-frequency components of image signals.

- An iterative scheme for joint estimation of motion parameters, the underlying HR pixel intensities as well as the parameters of the PAR model is proposed. In the second of our new VSR methods, estimating three groups of the above-mentioned unknown variables is a problem with a chicken-and-egg flavor. Therefore, the goal of the joint scheme is to achieve best statistical consistency among the PAR model parameters, motion parameters and the second-order statistics of reconstructed HR frames. In addition, this method estimates PAR model parameters iteratively from the reconstructed HR frames and hence overcomes the problem of PAR model mismatch. Also, it holds the possibility of mitigating motion estimation errors when using image interpolation to compute subpixel motion parameters.

1.5 Organization

The remaining of this thesis is structured into four chapters. In Chapter 2, we formulate the VSR reconstruction problem and outline the key related issues on this topic. In Chapter 3, we present a model-based interpolation scheme that learns the PAR model from observed data of multiple registered low-resolution video frames. In Chapter 4, we present a VSR method which combines the multi-frame observation model with the PAR model, where the PAR model plays a role of a regularization term. In this chapter, we also propose an iterative scheme for joint estimation of PAR model parameters, motion parameters as well as underlying high-resolution pixel intensities. In the end, we conclude this thesis with remaining challenges and future work for video super-resolution.

Chapter 2

Formulation of the VSR Problem

In this chapter, we investigate the VSR Problem and build a multi-frame observation model. With the observation model, VSR reconstruction is formulated as a least-squares problem. However, due to the lack of constraints, this least-squares problem is ill posed and consequently needs constraints to regularize its solution space. Then, we discuss the regularization issue for the ill-posed VSR problem. It is followed by a brief review of existing regularization methods. Next, we introduce a piecewise 2D autoregressive image model that plays a prominent role in our new algorithms in solving the ill-posed VSR problem. At the end of this chapter, we have an introduction to the issue of motion estimation in the VSR problem. It includes a brief review of most common motion estimation techniques, and a discussion on a variety of factors that decrease the accuracy of motion estimation.

2.1 Multi-frame Observation Model for Video Super-Resolution

The image resolution upconversion, namely the restoration of a single HR image from an observed (measured) LR version is a classic inverse problem. It can be linearly modeled as acquiring the LR image from the underlying clean HR image namely,

$$\boldsymbol{g} = \boldsymbol{D}\boldsymbol{H}\boldsymbol{z} + \boldsymbol{n} \tag{2.1}$$

where $\boldsymbol{z}$ is the underlying clean HR image of size $L_1N_1 \times L_2N_2$, and $\boldsymbol{g}$ is the observed degraded LR image of size $N_1 \times N_2$. Both vectors $\boldsymbol{z}$ and $\boldsymbol{g}$ denote the 2D images in a lexicographical (scanning) order. L_1 and L_2 are the down-sampling factors in horizontal and vertical directions respectively. Matrix $\boldsymbol{H}$ of size $L_1N_1L_2N_2 \times L_1N_1L_2N_2$ represents a low-pass filtering (i.e., blurring) operation which accounts for the optical point spread function (PSF) effects of the digital imaging system. Matrix $\boldsymbol{D}$ of size $N_1N_2 \times L_1N_1L_2N_2$ stands for the operation of decimation (or down-sampling). $\boldsymbol{n}$ is system random noise that normally is additive white Gaussian noise (AWGN). Specifically, $\boldsymbol{H}$, $\boldsymbol{D}$ and $\boldsymbol{n}$ are inherent to the camera system and independent of image signals.

In the scenario of reconstructing a video sequence of higher resolution, the linear image formation model in Eq. 2.1 is applicable to each of the underlying HR frames $\{\boldsymbol{z}_k\}$ and the associated observed LR frames $\{\boldsymbol{g}_k\}$. Here, symbol k denotes the time index of each frame. In addition, due to the temporal correlation of video frames, every two of the HR frame$\{\boldsymbol{z}_k\}$ and the LR frame $\{\boldsymbol{g}_l\}$ can be modeled as a sequence of geometry warping, blurring, decimation and corruption of additive noise namely,

$$\begin{aligned} \boldsymbol{g}_l &= \boldsymbol{D}\boldsymbol{H}_l\boldsymbol{G}(\boldsymbol{\nu}_l)\boldsymbol{z}_k + \boldsymbol{n}(k,l) \\ &= \boldsymbol{F}(l,\boldsymbol{\nu}_l)\boldsymbol{z}_k + \boldsymbol{n}(k,l) \end{aligned} \tag{2.2}$$

where matrix $\boldsymbol{G}(\boldsymbol{\nu}_l)$ of size $L_1N_1L_2N_2 \times L_1N_1L_2N_2$ denotes the geometry warping operation which is a function of motion parameter $\boldsymbol{\nu}_l$ between frame $\boldsymbol{z}_k$ and $\boldsymbol{z}_l$. Terms $\boldsymbol{H}$, $\boldsymbol{D}$ and $\boldsymbol{n}$ are the same as described above. In this formulation, the concatenation of $\boldsymbol{DHG}$ is simplified as matrix $\boldsymbol{F}$ of size $N_1N_2 \times L_1N_1L_2N_2$.

In consideration of temporal correlations between current frame $\boldsymbol{z}_k$ and its multiple neighbors, we can formulate the VSR problem by constructing a multi-frame observation model. As depicted in Fig. 2.1, the observation model represents a video frame formation process, namely that the underlying HR frame $\boldsymbol{z}_k$ yields multiple observed LR frames $\boldsymbol{g}_0, \cdots, \boldsymbol{g}_N$, via geometry warping, blurring, down-sampling and corruption of additive noise. Based on Eq. 2.2, the multi-frame observation model can be described by the following expression in a matrix-vector form.

$$\begin{bmatrix} \boldsymbol{g}_0 \\ \boldsymbol{g}_1 \\ \vdots \\ \boldsymbol{g}_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{F}(0,\boldsymbol{\nu}_0) \\ \boldsymbol{F}(1,\boldsymbol{\nu}_1) \\ \vdots \\ \boldsymbol{F}(N,\boldsymbol{\nu}_N) \end{bmatrix} \boldsymbol{z}_k + \begin{bmatrix} \boldsymbol{n}(k,0) \\ \boldsymbol{n}(k,1) \\ \vdots \\ \boldsymbol{n}(k,N) \end{bmatrix} \tag{2.3}$$

where there are N $(N \geq 1)$ observed LR frames, each of which probably consists of one or multiple local motions depending on the moving objects of the scene. Vector $\boldsymbol{\nu}_l$ $(l = 0, 1, \ldots, N)$ denotes one or a concatenation of multiple motion parameters between frame $\boldsymbol{z}_k$ and frame $\boldsymbol{z}_l$. Specifically, $l = 0$ refers to the current frame, i.e, vector $\boldsymbol{z}_k$, and thus $\boldsymbol{\nu}_0$ is a zero vector. Hereafter, we simplify the notations by defining the following expressions.

$$\begin{aligned} \boldsymbol{g} &= [\boldsymbol{g}_0^T, \boldsymbol{g}_1^T, \cdots, \boldsymbol{g}_N^T]^T \\ \boldsymbol{F}(\boldsymbol{\nu}) &= [\boldsymbol{F}^T(0, \boldsymbol{\nu}_0), \boldsymbol{F}^T(1, \boldsymbol{\nu}_1), \cdots, \boldsymbol{F}^T(N, \boldsymbol{\nu}_N)]^T \\ \boldsymbol{\nu} &= [\boldsymbol{\nu}_0^T, \boldsymbol{\nu}_1^T, \cdots, \boldsymbol{\nu}_N^T]^T \\ \boldsymbol{n} &= [\boldsymbol{n}^T(k, 0), \boldsymbol{n}^T(k, 1), \cdots, \boldsymbol{n}^T(k, N)]^T \end{aligned} \tag{2.4}$$

Then, the multi-frame observation model in Eq. 2.3 can be simplified into a linear form as

$$\boldsymbol{g} = \boldsymbol{F}(\boldsymbol{\nu})\boldsymbol{z} + \boldsymbol{n} \tag{2.5}$$

2.2 Regularization for Video Super-Resolution

As addressed in the preceding section, the reconstruction of high-resolution image(s) from a single or collection of observed images is an inverse problem. With the image formation model in Eq. 2.1, solving this restoration problem can be formulated as minimizing the effects of system noise $\boldsymbol{n}$. It assures certain fidelity of the final solution to the observed data on condition that the system matrix $\boldsymbol{H}$ and $\boldsymbol{D}$ namely, the PSF function of the imaging system and the upconversion factor are known. From a statistical perspective, the optimal solution in least-squares (LS) sense is attained by

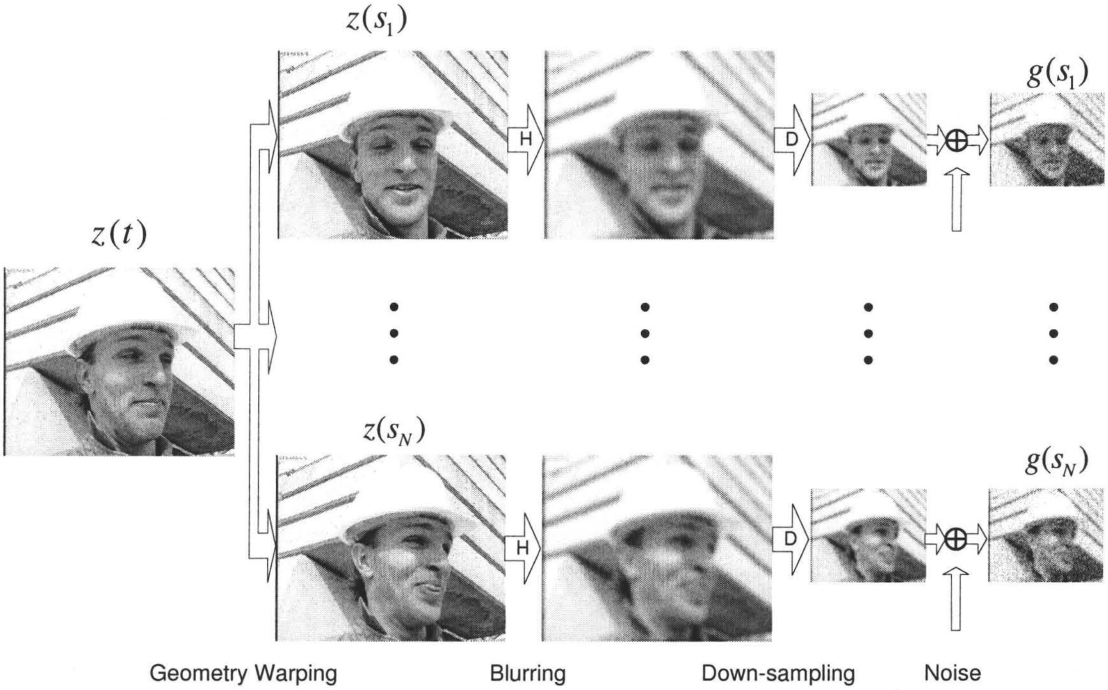


Figure 2.1: Illustration of the multi-frame observation model

minimizing the L_2 norm of noise $\boldsymbol{n}$ namely,

$$\hat{\boldsymbol{z}} = \arg\min_{\boldsymbol{z}} \left\| \boldsymbol{g} - \boldsymbol{D}\boldsymbol{H}\boldsymbol{z} \right\|_2^2 \tag{2.6}$$

When image resolution is upconverted by a factor of κ ($\kappa > 1$), the size of vector $\boldsymbol{z}$ is κ times larger than that of vector $\boldsymbol{g}$. As such the number of observations is smaller than that of the unknowns, which indicates that estimating $\boldsymbol{z}$ is always an ill-posed problem.

In the VSR problem, the number of observations is increased thanks to the availability of multiple low-resolution frames. Based on the multi-frame observation model in Eq. 2.5, the restoration of HR frame(s) from observed LR frames can be formulated as solving the following optimization problem.

$$\hat{\boldsymbol{z}} = \arg \min_{\boldsymbol{z}} \left\| \boldsymbol{g} - \boldsymbol{F}(\boldsymbol{\nu})\boldsymbol{z} \right\|_2^2 \tag{2.7}$$

This optimization process attempts to minimize the effects of system noise $\boldsymbol{n}$ in terms of L_2 norm. The solution $\hat{\boldsymbol{z}}$ is the maximum likelihood (ML) estimate of $\boldsymbol{z}$ if noise $\boldsymbol{n}$ is zero-mean AWGN[7].

Algebraically, the optimal solution in least-squares sense for the optimization problem in Eq. 2.7 can be written as

$$\hat{\boldsymbol{z}} = (\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{g} \tag{2.8}$$

on the condition that matrix $\boldsymbol{F}$ is known and matrix $\boldsymbol{F}^T\boldsymbol{F}$ is non-singular. Therefore, computing the solution $\hat{\boldsymbol{z}}$ requires inverting square matrix $\boldsymbol{F}^T\boldsymbol{F}$. This makes the solution $\hat{\boldsymbol{z}}$ highly sensitive to the condition number of matrix $\boldsymbol{F}^T\boldsymbol{F}$. If matrix $\boldsymbol{F}^T\boldsymbol{F}$ is well conditioned, solution $\hat{\boldsymbol{z}}$ can be directly computed. However, more often than not, matrix $\boldsymbol{F}^T\boldsymbol{F}$ is ill conditioned.

As presented in Section 2.1, matrix $\boldsymbol{F}$ concatenates geometric warping operation $\boldsymbol{G}$, blurring operation $\boldsymbol{H}$ and down-sampling operation $\boldsymbol{D}$. Among these operations, the blurring operation $\boldsymbol{H}$ can be estimated from the digital imaging system[1]. In practice, it is more common to determinate $\boldsymbol{H}$ by assuming a linear space-invariant PSF function for the imaging system. The down-sampling operation $\boldsymbol{D}$ solely depends on the upconversion factor κ. Matrix $\boldsymbol{G}$ involves geometry warps of temporally correlated pixels. Therefore, matrix $\boldsymbol{F}$ relies on the registration of observed low-resolution video frames. Historically, most VSR algorithms solve the optimization problem in Eq. 2.7 in two steps[3]:

1. motion estimation (or image registration) to determinate matrix $\boldsymbol{F}$;
2. solving an inverse problem based on the first step.

Motion estimation performed in the first step simplifies the optimization problem as a linear least-squares problem with only the underlying high-resolution pixel intensity $\boldsymbol{z}$ unknown.

Even though matrix $\boldsymbol{F}$ can be determined through explicit motion estimation, it is in fact spare and sensitive to the adopted motion estimator (namely its accuracy). In this case, if matrix $\boldsymbol{F}^T\boldsymbol{F}$ is ill conditioned, inverting matrix $\boldsymbol{F}^T\boldsymbol{F}$ would amplify the effects of motion estimation error and consequently lead to significant reduction in the performance of the final solution. Moreover, invertibility of matrix $\boldsymbol{F}^T\boldsymbol{F}$ depends on the number of available LR frames(i.e, independent observed pixels). If matrix

$\boldsymbol{F}^T\boldsymbol{F}$ is not invertible (i.e., singular), the optimization problem would become worse, as it is underdetermined (i.e., ill posed).

Under this circumstance, it is necessary to impose a regularization term $\rho(\boldsymbol{z})$ as a constraint on the solution space of the ill-posed inverse problem. As such the constraint can stabilize this ill-posed problem and make its solution more accurate by solving the following Lagrangian[3].

$$\hat{\boldsymbol{z}} = \arg \min_{\boldsymbol{z}} \left\{ \left\| \boldsymbol{g} - \boldsymbol{F}(\boldsymbol{\nu})\boldsymbol{z} \right\|_2^2 + \lambda\rho(\boldsymbol{z}) \right\} \tag{2.9}$$

where Lagrange Multiplier λ adjustes the strength of the regularization term $\rho(\boldsymbol{z})$, and hence provides a balance between the fidelity term $\left\| \boldsymbol{g} - \boldsymbol{F}(\boldsymbol{\nu})\boldsymbol{z} \right\|_2^2$ and the regularization term. In practice, the value of λ is commonly chosen based on visual quality of the VSR reconstructed results.

In general, regularization term $\rho(\boldsymbol{z})$ incorporates some prior knowledge extracted from the observed low-resolution image(or video frame). This knowledge can be local pixel structures of the image, such as edge gradient, smoothness, etc. By using the prior knowledge, the regularization term $\rho(\boldsymbol{z})$ preserves certain coherence of pixel structures between the observed low-resolution image and the underlying high-resolution image. The most common regularization method used in VSR is total variation (TV) [7, 9, 14, 30]. In general, it can be described by the following expression.

$$\rho(\boldsymbol{z}) = TV(\boldsymbol{z}) = \underbrace{\sum_{l=-p}^{p} \sum_{m=0}^{p}}_{l+m\geq 0} \alpha^{|m|+|l|} \left\| \boldsymbol{z} - S_x^l S_y^m \boldsymbol{z} \right\|_1 \tag{2.10}$$

where operator S_x^l and S_y^m represent shifts of image pixel $\boldsymbol{z}$ by l and m pixel units in horizontal and vertical directions respectively. Term α $(0 < \alpha < 1)$ is a constant which spatially adjusts the effects of differentials. TV-based image restoration methods use L_1 norm of the magnitude of image gradient to regularize deblurring. They reconstruct the underlying high-resolution image by solving nonlinear partial differential equations (PDE) with the gradient constraints. Even though TV methods have been shown to be effective in reproducing large-scale edges[32], there are two drawbacks that make TV incapable of preserving small image details. First, proper norm of TV in image restoration is L_1[32]. It works based on an idealistic assumption that the first-order derivative of the image signal keeps constant in a small-scale area, and the higher order derivatives are all zero valued. As a matter of fact, for both natural images and computer synthesized images, there is a great high-order (especially second-order) statistical abundance of image signal waveform that accounts for subtle image details. Second, the lack of adaptivity in adjusting its own parameters makes TV incapable of spatially varying in multi-scale image features e.g., large-scale edges, and fine texture details[15]. In other words, it does not distinguish small details from large-scale edges/textures over an image. Due to these properties, TV-based image restoration methods can not preserve image details but force piecewise smoothness of image signals especially in the presence of image noise[5].

Instead of utilizing TV methods, we propose model-based approaches to regularize the solutions for the ill-posed inverse problem in this thesis. Compared with TV methods, our approaches are based on a piecewise 2D autoregressive image model, and are more capable of representing image waveforms ranging from smooth shades, periodic textures to transients like edges. Furthermore, they can adapt the reconstruction of high-resolution images to local varying image waveforms. Therefore, our model-based approaches can gain superior performance against TV-based counterparts for VSR reconstruction.

2.3 An Introduction to Piecewise 2D Autoregressive Image Model

During the past decade, image modeling has been a challenging research topic in imaging processing areas, such as image compression, image restoration etc. For both natural and computer synthesized images, the structure of local image waveform varies spatially over the image, which results in the non-stationarity of the second-order statistics of image signal. Therefore, modeling of the non-stationary image waveform needs to be highly adaptive to the varying local pixel structures. In the end of 1990s, Wu *et al.* had a measured success in a research on predictive lossless image compression[41, 42]. In that work, image signal is modeled as a piecewise 2D autoregressive (PAR) process on the assumption of piecewise stationarity of image signals. The model parameters are adaptively estimated from pixel samples of a moving local window on a pixel-by-pixel basis across the image. In light of the predictive coding of pixels in that work, the autoregressive model is designed to be causal to the current pixel. Later, Wu and Zhang proposed to apply the PAR model to still image interpolation[44]. They assume that image signal preserves the spatial coherence of pixel structures regardless the change of image resolution. Thus, PAR model parameters are estimated for each pixel of the observed low-resolution image and then applied to fit the underlying high-resolution image. Additionally, the PAR model in still image interpolation is not causal. Compared with the causal PAR counterpart, it takes more advantage of sample statistics of the local window.

The advantage of the PAR model lies in two aspects when applied to still image interpolation. First, the adaptive learning of local pixel structures from low-resolution image is performed by taking into consideration the statistics of a local window (or a block) instead of pixels in isolation. Second, rather than individual estimation of each missing high-resolution pixel, a block of missing pixels in relation to the nearby known pixels are simultaneously estimated by fitting them to the learnt PAR model. Therefore, the blockwise estimation of missing pixels ensures certain spatial coherence of the reconstructed image. Due to the advantages above, the PAR model provides competitive solutions for adaptive still image interpolation[44] and some related applications[45].

Despite these advantages, the PAR model suffers two shortcomings. First, the model parameters are estimated from the low-resolution still image (or video frame) but applied to reconstruct the high-resolution image in still image interpolation. With the change of image resolution, the spatial correlation of pixels varies for different scaling (i.e, pixel distance). This potentially leads to an inconsistency of the second-order statistics of image signals, and consequently incurs a mismatch of the PAR model between the low-resolution and high-resolution image. Second, it is likely that the blockwise estimation of missing high-resolution pixels encounters a dilemma of model overfitting. Mathematically, to solve an array of equations, the number of independent observations is required to meet the number of unknown variables. In other words, the number of observed pixels needs to be large enough for a robust estimate of the PAR model. However, owing to the piecewise stationary nature of image signals, a 2D AR model holds only within a small local window and therefore insufficient observation data can be provided. In [44], the limitation of piecewise stationarity of image signals against the minimum required number of observations is well balanced by choosing a low-order PAR model and a moderate-sized window.

In this thesis, we advocate the use of the PAR model to regularize the underdetermined high-resolution image signals in the VSR problem. In details, two new methods are proposed in the coming two chapters to solve the above-analyzed dilemma of learning PAR models from observed low-resolution images.

2.4 Motion Estimation in Video Super-Resolution

In VSR reconstruction, it is necessary to register pixels of observed LR frames onto an HR image grid at subpixel accuracy, such that each of the LR frames can contribute substantial pixel information to the reconstruction of an HR frame. In general, the frame whose HR version is to be reconstructed is known as a current frame. The observed neighboring frames are registered against the current frame. As illustrated in Fig. 1.3, this image registration procedure, in essence, is estimating the subpixel displacements between the LR frames. In this sense, image registration and the displacements are also known as motion estimation and motion parameters(or motion vectors) respectively. The accuracy of estimated motion parameters is critical to the performance of observation model based VSR algorithms[33]. In practice, inaccurate or incorrect motion parameters can result in visually noticeable artifacts on the reconstructed high-resolution frame, and hence have a disastrous influence over the performance of super-resolution algorithms. While, the accurate estimation of arbitrary motions in a natural video sequence is an extremely difficult task, and the performance of estimators can not be guaranteed. For a scene in the video sequence, it may contain a global or multiple local motions, and these motions can be translational, rotational, zooming or even a combination of the above all. In light of the above, motion models and estimation methods in VSR should be appropriately chosen in accordance with the *a priori* knowledge of motions in the video sequence.

The most popular motion estimators in the literature can be classified into two categories: feature-based and area-based motion estimation. Feature-based methods take the advantage of image features such as edges, points, and line intersections etc. Compared with area-based methods, feature-based methods are much more robust against image noise and image degradation. But the disadvantages are manifest. As image features are sparsely distributed, not all of pixels in an image can be registered. In contrast to feature-based methods, area-based methods take into consideration all pixels aside from feature pixels over the image. For this reason, the computational cost of these methods is prohibitively high.

In terms of camera motion, motion estimators can be categorized into two classes: parametric and nonparametric methods. The parametric methods include 4 or 6-parameter affine model, and 8-parameter projection model[25] etc. They take advantage of 2D parametric transformations (e.g., 2D affine, 2D quadratic and 2D projective), and attempt to assign a parametric motion model to each group of pixels that have an identical camera motion. Thus, they can achieve high computational efficiency when dealing with global motions. If multiple local motions exist in a scene, parametric methods first need to isolate each of the local moving objects. Afterwards, one parametric model for each object is computed through what often is an iterative minimizing process. In contrast to parametric motion estimators, nonparametric methods estimate motions on a pixel-by-pixel or block-by-block basis. They do not group pixels in terms of identical camera motions. Therefore, nonparametric methods can achieve high computational efficiency when dealing with a scene with multiple local motions.

Among the nonparametric motion estimators, Block Matching Algorithm (BMA) is a block-based method for locating matched blocks or identifying motion parameters between two images/frames[26]. It searches within a local window under an idealistic assumption that motion field within a block is uniform. Due to its simplicity in implementation, BMA has been widely used in visual tracking, video surveillance and video compression standards, e.g., MPEG-1, MPEG-2, MPEG-4, H.263, H.264 etc.

In a general situation, objects within adjacent frames move at certain finite displacements, which results in pixel correlation between the adjacent frames. The pixel correlation is also known as temporal correlation of video frames. In this case, a block in a current frame can be highly temporally correlated with its peers in an adjacent (or neighboring) frame. Among these temporally correlated blocks, the two that carry the minimum difference are known as the best matched blocks. Correspondingly, the displacement between them is referred as motion parameter (dx, dy), where dx and dy are horizontal and vertical displacements respectively. In practice, BMA can be used to identify motion parameters for a single pixel, or non-overlapping blocks in which all pixels presumptively have identical motion parameters.

On the quest to attain subpixel level motion parameters, motion estimators always take advantage of an image interpolation technique, such as bilinear, bicubic interpolation method[18] etc. However, the computation of such motion parameters (e.g., half-pixel, quarter-pixel, and 1/8-pixel) using an exhaustive search (or full search) strategy is very intensive. During the past decade, much research attention has been paid in this regard to improve the efficiency of BMA. Among the fast BMA methods are three-step search (3SS)[22], cross-search [11], diamond search[46], new three-step search (NTSS) [23], four-step search(4SS) [29], block-based gradient descent search (BBGDS)[24]. These fast BMA algorithms save the computational cost by means of certain search patterns and the reduced number of searching points. However, the computational efficiency is achieved at the cost of lower accuracy.

Aside from reducing the number of searching points, there are still some other factors that decrease the accuracy of BMA algorithms. For example, image noise and image degradation is inevitable for images acquired through image acquisition devices. It lowers the accuracy of motion estimation by contaminating the true values of pixel intensities. In addition, interpolation errors are unavoidable either. As image interpolation is an ill-posed inverse problem, it can make the interpolation value approach the true pixel intensity, but can hardly guarantee a result that is free of interpolation errors.

Chapter 3

Multi-frame based PAR Model for Video Super-Resolution

As mentioned in Section 2.3, the PAR model faces a dilemma when applied to image interpolation. On one hand, correct reproduction of missing high-resolution pixels relies on a valid model of the underlying image signal; On the other hand, model of the image signal can be built only if the image signal is available. In still image interpolation, the PAR model is learned from an observed LR image and applied to reconstruct the underlying HR image[44]. However, if a mismatch between the second-order statistics of the LR and HR image occurs, the PAR model learned from the LR image would not fit the underlying HR image, which potentially leads to poor interpolation performance. One way to resolve this dilemma is to increase pixel density (i.e., spatial resolution) of the observed pixel samples, such that the mismatch of the second-order statistics can be mitigated. In still image interpolation, this approach can not be easily applied due to an insufficient number of observed pixel samples. By contrast, the abundance of temporal correlations of video frames provides the possibility to resolve this dilemma.

As analyzed in Section 2.4, pixels of multiple observed low-resolution video frames at subpixel displacements can be registered onto a high-resolution grid. These pixels are true samples that originate in the underlying high-resolution image. Therefore, if registered pixels on the high-resolution grid are distributed at regular pixel sites as illustrated in Fig. 1.3, then they can be treated as good estimates of the underlying high-resolution pixels. In this case, the PAR model learned from the registered pixel samples on the high-resolution grid would be much more accurate than that learned from the low-resolution pixel samples, which reduces the likelihood of model mismatch. In addition, since the number of observations in a local window increases without violating the piecewise stationary nature of image signals, this method reduces the possibility of model overfitting (as analyzed in Section 2.3).

In addition to the problem of PAR model learning, another problem addressed in this chapter is motion estimation and the synthesis of high-resolution blocks (i.e., local windows). The role of motion estimation is to register observed pixels of multiple neighboring low-resolution video frames, such that high-resolution local windows can be built for PAR model learning. Many available motion estimators used in VSR, for example the one in [37], try to describe a global or local motion via a 4-parameter or 6-parameter affine motion model at subpixel precision. Computing the affine motion is prohibitively expensive, as it often involves an iterative procedure to solve a minimization problem[3]. In the case of a scene with multiple local motions, motion estimators in the literature often try to perform motion segmentation before computing each of them. This operation makes the motion estimation problem even more complicated[36].

Another advantage of our new VSR method is the use of the simplest Block-Matching Algorithm (BMA) for motion estimation. As reported in the literature, BMA is computationally much less intensive than parametric motion estimators. In addition, both BMA and the PAR model are block-based methods. The natural integration of BMA and the PAR model makes the block-based approaches capable of dealing with both global and local motions. Simulations have been conducted on natural video sequences and the results convincingly demonstrate the improved performance of the PAR model in spatial resolution upconversion.

This rest of this chapter is organized as follows. Section 3.1 briefly describes the piecewise autoregressive image model. Section 3.2 discusses the block-based motion estimation with an aim to register pixels of multiple observed low-resolution frames onto a high-resolution image grid. Section 3.4 presents a model-learning scheme such that PAR model parameters can be estimated from pixel samples on the high-resolution grid. This section also discusses the reconstruction of missing high-resolution pixels. Simulation results and discussion are presented in Section 3.5.

3.1 Learning of Piecewise 2D Autoregressive Model

It is observed that pixels in a local window can be modeled by a piecewise 2D autoregressive process. This process can be described by the following expression[44]:

$$I(i,j) = \sum_{(m,n)\in W} \alpha_{m,n} I(i+m, j+n) + \nu_{i,j} \qquad (3.1)$$

where pixel $I(i,j)$ is predicted by its neighbors $I(i+m, j+n)$ in a local window W. Term $\nu_{i,j}$ is a random perturbation independent of pixel site (i,j) and pixel intensity $I(i, j)$. It accounts for both the fine-scale randomness of image signal and measurement noise.

Model parameter $\alpha_{m,n}$ specifies the structure, namely the direction and amplitude of features within the local window $W_{i,j}$. In fact, edge directions of natural images are randomly distributed. Therefore, the order of the PAR model should be appropriately chosen to achieve the best model fitting[2]. In this thesis, we build two separate 4-parameter PAR models considering general cases. These two PAR models fit two sets of pixel samples, namely 8-connected neighborhood and 4-connected neighborhood of pixel intensity $x_i \in W$. As depicted in Fig. 3.1, the two models are specified by two groups of autoregressive coefficients $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$. They characterize the axial and diagonal correlation respectively. More details about the PAR model can be referred from [44].

Natural image signal may not be stationary in a large scale due to its spatially inconsistent second-order statistics. Nevertheless, edges and textures in forms of contiguous pixels tend to manifest consistent spatial characteristics in a small scale. It suggests that PAR model parameters in the locality remain constant or near constant. Therefore, it is reasonable and acceptable to assume the piecewise stationarity of image signal. This forms the fundamental assumption on which the autoregressive image model works. Mathematically, the structure of image signals within the local window can be learnt by fitting the autoregressive model to pixels in the local window W. This fitting process is formulated as solving the following two least-squares problems:

$$\begin{aligned}\hat{\boldsymbol{\alpha}} &= \arg\min_{\boldsymbol{\alpha}}\left\{\sum_{i\in W}(y_i - \sum_{0\le k\le 3}\alpha_k y_{i,k}^{\times})^2\right\}\\ \hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}}\left\{\sum_{i\in W}(y_i - \sum_{0\le k\le 3}\beta_k y_{i,k}^{+})^2\right\}\end{aligned} \tag{3.2}$$

where PAR parameter $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ in least-squares sense account for an optimal solution for the corresponding PAR model within local window W.

To distinguish pixels of a low-resolution frame and a high-resolution frame, hereafter we have $\boldsymbol{y}$ and $\boldsymbol{x}$ denote pixel intensities in the observed low-resolution frame and the underlying high-resolution frame respectively.

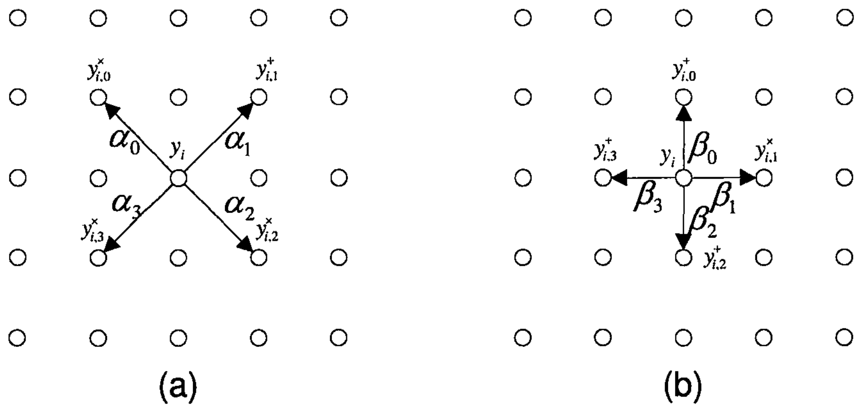
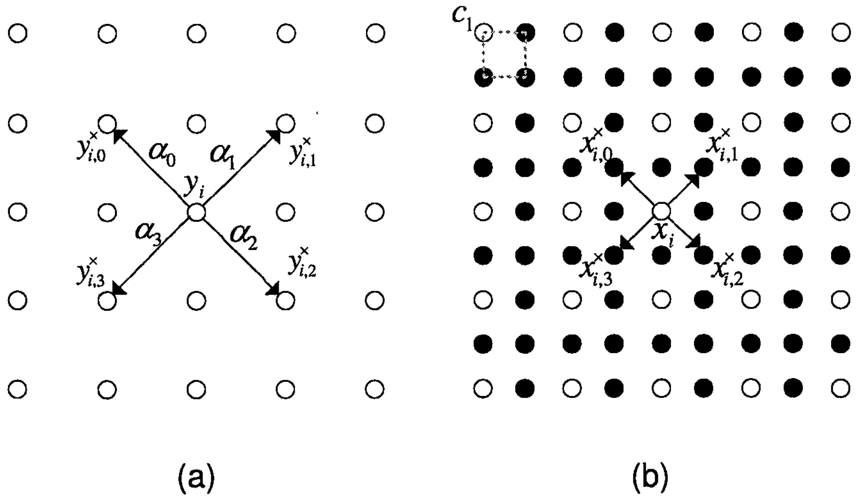


Figure 3.1: PAR model: (a) diagonal mode, (b) axial mode.

3.2 Block Matching

As analyzed at the beginning of this chapter, the possibility of PAR model mismatch between the LR video frame and the underlying HR video frame can be reduced by increasing pixel density of local windows in the observed video frame. In practice, the pixel density can be increased by registering pixel samples from multiple LR reference frames. In this thesis, the registration or motion estimation, is performed by means of block matching (BMA). This method finds the best matched block within a local search window W_S, using a matching criterion of Euclidean norm:



$\overset{\times}{y}_{i,0}$
α_0
α_1
$\overset{\times}{y}_{i,1}$
y_i
α_3
α_2
$\overset{\times}{y}_{i,3}$
$\overset{\times}{y}_{i,2}$
c_1
$\overset{\times}{x}_{i,0}$
$\overset{\times}{x}_{i,1}$
x_i
$\overset{\times}{x}_{i,3}$
$\overset{\times}{x}_{i,2}$
(a)
(b)

Figure 3.2: Comparison of Learning PAR model from a LR image (a) with from an HR image (b). White dots represent observed pixels; black dots represent interpolated pixels.

$$D_n(dx, dy) = \sum_{i}^{M}\sum_{j}^{N} |I_c(i,j) - I_r(i + dx, j + dy)|^n \tag{3.3}$$

where M, N are the height and width of a registration block. $I_c(\cdot, \cdot)$ and $I_r(\cdot, \cdot)$ represent pixels in current frame and reference frame respectively. Term n is a positive integer. In practice, the 1-norm of the matching error $D_1(dx, dy)$ is commonly used. Vector (dx, dy) defines the translational displacement between the current block and the reference block. Then the displacement of the best matched block with respect to the current block is given by

$$(\hat{dx}, \hat{dy}) = \arg \min_{(dx,dy)\in W_S} D_1(dx, dy) \tag{3.4}$$

In the VSR problem, the image resolution upconversion factor κ can be an arbitrary positive fractional number larger than one. In the proceeding discussion, we restrict factor κ to be a power of two to simplify the motion estimation problem. Then for an integer κ e.g., 2, 4, etc., block matching is performed at $1/\kappa$ precision e.g., half-pixel or quarter-pixel precision etc. For the clarity of our presentation and without loss of generality, we develop our VSR method for upconversion factor of $\kappa = 2$. In this case, half-pixel precision BMA is needed.

By performing a half-pixel precision BMA on the current low-resolution frame and its neighbors, multiple registered low-resolution blocks can form a new block whose resolution is twice that of the low-resolution frame. In the resulting high-resolution block, current low-resolution block and multiple registered reference blocks have integral displacements $(\hat{dx}, \hat{dy})$. In terms of parity, the displacements (dx, dy) fall into four classes: $(\mathcal{O}, \mathcal{O})$, $(\mathcal{O}, \mathcal{E})$, $(\mathcal{E}, \mathcal{O})$ and $(\mathcal{E}, \mathcal{E})$. Here, the decorated letters $\mathcal{O}$ and $\mathcal{E}$ denote an odd-valued and even-valued displacement value respectively. The four classes are illustrated in Fig. 3.3. In the following section, we will discuss on synthesis of multiple registered low-resolution blocks to form a high-resolution block.

3.3 Synthesis of Multiple Low-resolution Blocks

First of all, let us consider the case as depicted in Fig. 3.3(a). Two blocks of low-resolution pixels denoted by black dots from reference frame and white dots from current frame respectively, are registered onto a high-resolution grid. The block of black dots (denoted by B_r) have an integral spatial shift $(\mathcal{O}, \mathcal{O})$ to the block of white dots (denoted by B_c). In this case, a new high-resolution block B_h can be synthesized by multiplexing the two blocks. In view of block-matching error, the synthesis is performed by using a weighted fusing scheme instead of simple multiplexing.

Since the two blocks B_c and B_r at an integral displacement $(\mathcal{O}, \mathcal{O})$ are spatially interleaved, one can be estimated from anther through image interpolation. Suppose that a pixel $\hat{y}_i$ in block B_r corresponds to a pixel y_i in the high-resolution block B_h, i.e., the black dots. Its another estimate $\hat{y}_i^*$ is interpolated from current block B_c. Then, the fusing of $\hat{y}_i$ and $\hat{y}_i^*$ yields a more robust estimate of y_i:

$$y_i = w\hat{y}_i + (1 - w)\hat{y}_i^* \tag{3.5}$$

where w $(0 \leq w \leq 1)$ is a context-based weight determined by the matching degree $d_i = |\hat{y}_i - \hat{y}_i^*|$ at each pixel site within the block.

$$w = \begin{cases} w_1 & \text{if} \quad d_i \leq 2 \\ w_2 & \text{if} \quad 2 < d_i \leq T \\ w_3 & \text{if} \quad T \leq d_i \end{cases} \tag{3.6}$$

where T is a threshold optimized through simulations.

As for the registration cases with displacements $(\mathcal{O}, \mathcal{E})$ and $(\mathcal{E}, \mathcal{O})$, the principle of synthesizing a high-resolution block is fundamentally the same as the case $(\mathcal{O}, \mathcal{O})$ discussed above. The only difference lies in the pixels modified in block B_h, namely the black dots shown in Fig. 3.3(b)(c). One thing worth noting is that pixels (i.e,

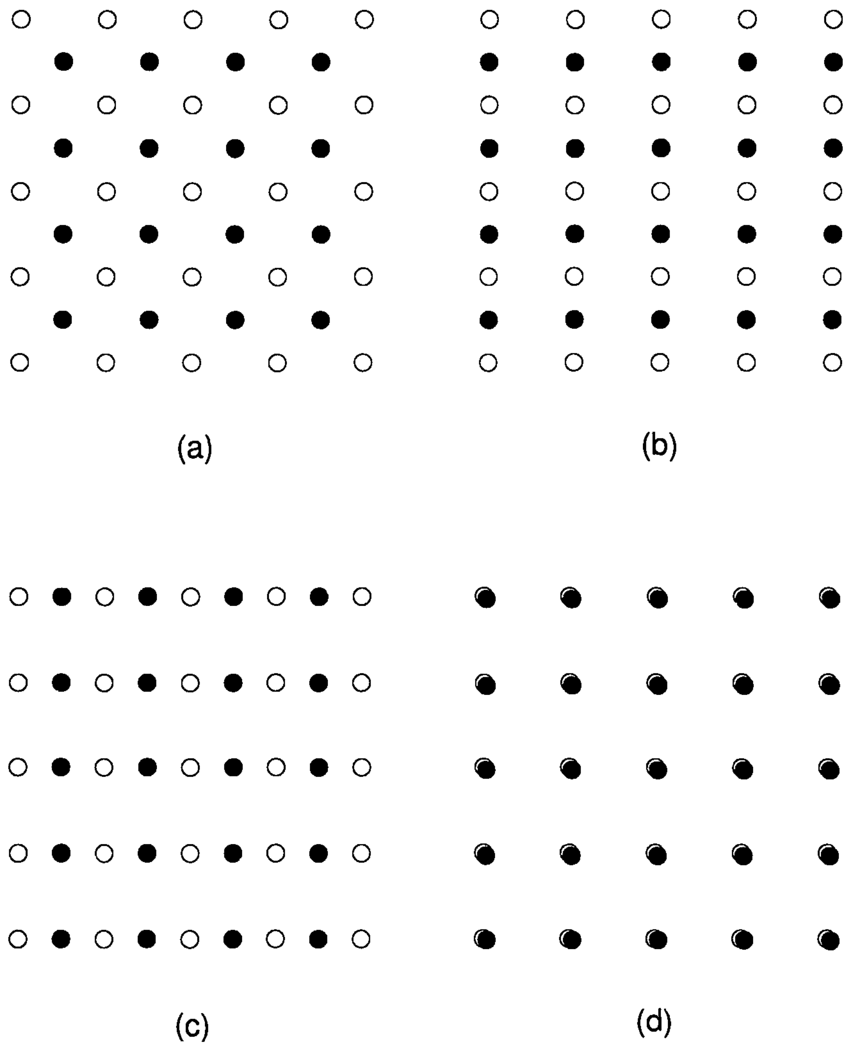


Figure 3.3: Synthesis of two low-resolution blocks at displacement: (a) $(\mathcal{O},\mathcal{O})$; (b) $(\mathcal{E},\mathcal{O})$; (c) $(\mathcal{O},\mathcal{E})$; (d) $(\mathcal{E},\mathcal{E})$. Black dots represent registered pixels from a reference frame; white dots represent pixels registered from the current frame.

white dots) in block B_h corresponding to those in block B_c originate in the underlying high-resolution current frame. Therefore, they are never altered in light of their reliability. For the case with displacement $(\mathcal{E}, \mathcal{E})$, registered pixels from neighboring frames contain the same pixel information as those from the current frame. Therefore, they can not help increase pixel density of the high-resolution grid (i.e., block) and hence are discarded.

The goal of image registration in this method is to increase the density of observed pixel samples such that it can approach that of the underlying high-resolution pixels. In the case of upconverting video frames of size $N_1 \times N_2$ by a factor of κ ($\kappa > 1$), there are $\kappa^2 N_1 N_2$ pixel sites on the high-resolution image grid. Since the pixels of the current frame (as denoted by white dots in Fig. 3.2) occupy $N_1 N_2$ of those sites, there are still $(\kappa^2 - 1)N_1 N_2$ remaining pixel sites to be filled by multiple neighboring LR frames. In the case of $\kappa = 2$ as shown by a dashed rectangle in Fig. 3.2, there are $\kappa^2 - 1 = 3$ unoccupied sites (denoted by black dots) around each white dot to be filled by pixels of neighboring LR frames. In what follows, we will study the probability that all the unoccupied pixel sites are filled by pixels of r registered neighboring LR frames ($r > 1$).

The problem of filling pixel sites as stated above can be formulated as using r pixels to fill n sites ($n = \kappa^2$), one of which (denoted as c_1) has already been filled by a pixel of the current frame. Due to arbitrary motions of moving objects or/and cameras, a registered pixel of a reference frame has equal probability $\frac{1}{n}$ to fill one of the n pixel sites. First, let us consider the event A_m that r pixels fill $n-m$ unoccupied sites, leaving m sites unfilled. The probability of event A_m is given by [10],

$$P_m(r,n) = \begin{pmatrix} n \\ m \end{pmatrix} \sum_{i=0}^{n-m} (-1)^i \begin{pmatrix} n-m \\ i \end{pmatrix} \left(1 - \frac{m+i}{n}\right)^r \tag{3.7}$$

However, if pixel site c_1 has already been filled beforehand, then the m ultimately unfilled sites should only be from the remaining $n-1$ ones. Let B be the event that there are $r-i$ pixels ($i = 0, 1, \ldots, r$) falling into the remaining $n-1$ sites. These $r-i$ pixels can be chosen in $\begin{pmatrix} r \\ r-i \end{pmatrix}$ different ways, and for the given $r-i$ pixels, they can fill the $n-1$ sites in $(n-1)^{(r-i)}$ different ways. Thus, there are $\begin{pmatrix} r \\ r-i \end{pmatrix} (n-1)^{(r-i)}$ ways of choosing $r-i$ pixels from r ones to fill $n-1$ sites. Since the total number of possible arrangements in filling n sites with r pixels is n^r, the probability of event B therefore is $\begin{pmatrix} r \\ r-i \end{pmatrix} (n-1)^{(r-i)}/n^r$. Then the probability of event A_m namely that the $r-i$ pixels leave m of the $n-1$ sites unfilled is $\begin{pmatrix} r \\ r-i \end{pmatrix} (n-1)^{(r-i)}/n^r \cdot P_m(r-i, n-1)$. Accordingly, the total probability of event A_m is the summation of those for all i's namely,

$$P'_m(r,n) = \sum_{i=0}^{r} \begin{pmatrix} r \\ r-i \end{pmatrix} (n-1)^{(r-i)}/n^r \cdot P_m(r-i, n-1) \tag{3.8}$$

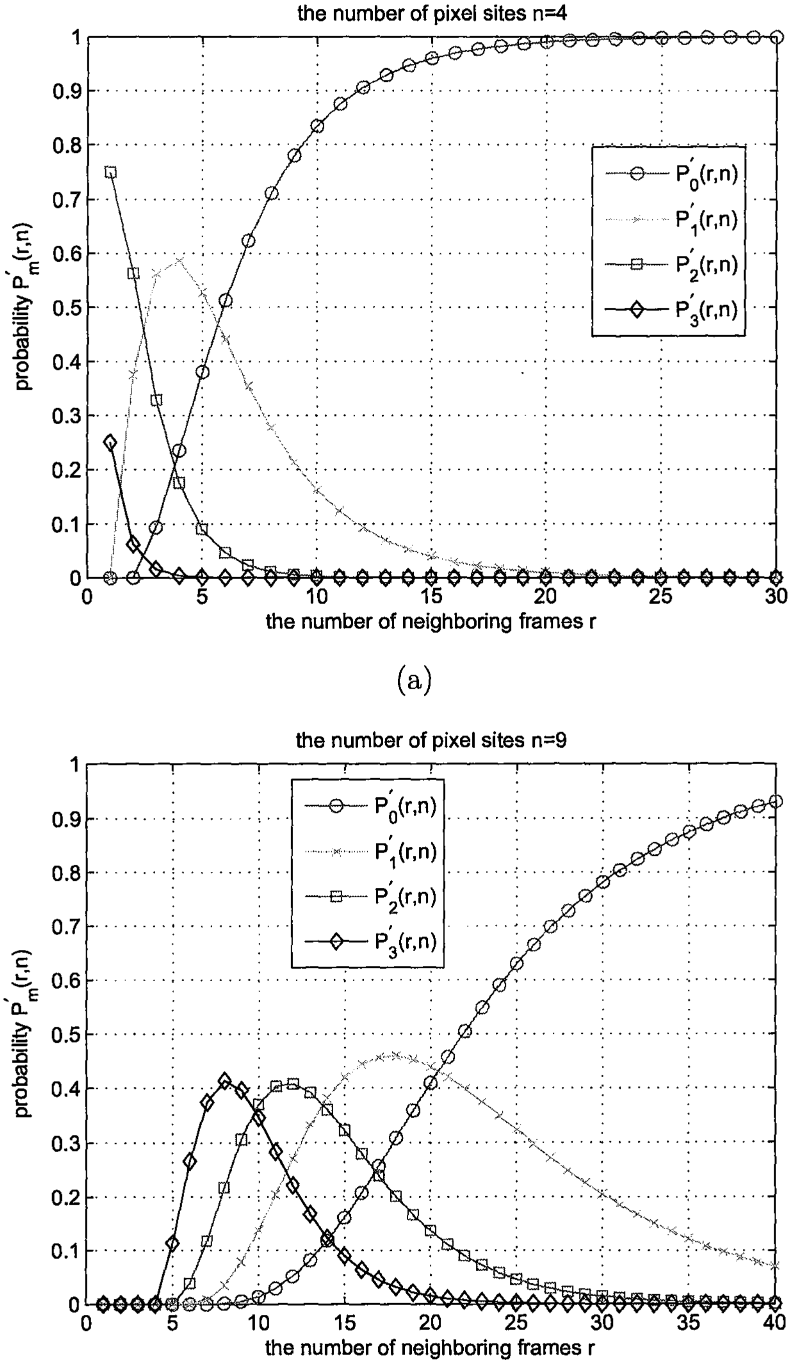
In our case, we use $m = 0$ to get all pixel sites filled and the corresponding probability is $P'_0(r,n)$. Moreover, we also have the probability distributions for cases that there are 1, 2, 3 pixel sites unfilled, namely $P'_1(r,n)$, $P'_2(r,n)$ and $P'_3(r,n)$. Fig. 3.4

illustrates $n = 4, 9$ and corresponding probability $P'_m(r, n)$ versus the number of registered neighboring frames r. From this figure, we can find that probability $P'_0(r, n)$ approaches to 1, while the other probabilities decline to zeros as the number of neighboring frames r increases. Therefore, in the case of upconversion factor $\kappa = 2$ (i.e., $n = 4$), it is confident to conclude that the unoccupied pixel sites can be fully filled at probability higher than 0.8 when the number of neighboring frames is larger than 10.

With the registration of multiple low-resolution blocks, a high-resolution block can be produced at relatively high probability (i.e., larger than 0.8) that all pixel sites are occupied. Estimating of PAR model parameters in the high-resolution block(i.e., local window) is performed by following the model-learning scheme stated in Section 3.1. However, it is possible to encounter the cases that pixel sites on the high-resolution block are not fully filled, such as those illustrated in Fig. 3.3(a)(b)(c). In these cases, pixel distances within the PAR model are inconsistent in various directions. In our method, we do not fill the unoccupied pixel sites via image interpolation, for interpolated pixels at low accuracy can not improve estimates of PAR model parameters. Instead, we use low-resolution blocks of the current frame for model learning.

3.4 Reconstruction of High-Resolution Pixels via PAR model

On the basis of the PAR model learned in a high-resolution local window, the reconstruction of the underlying high-resolution pixel intensity values x can be formulated



(a)

(b)

Figure 3.4: Probability distributions for upconversion factor $\kappa = 2, 3$.

as minimizing the following least-squares inverse filtering problem.

$$\min_{x}\left\{\xi^{\times}\sum_{i\in W}(x_i-\sum_{0\le k\le 3}\hat{\alpha}_k x_{i,k}^{\times})^2+\xi^{+}\sum_{i\in W}(x_i-\sum_{0\le k\le 3}\hat{\beta}_k x_{i,k}^{+})^2+\lambda\left\|\boldsymbol{x}*\boldsymbol{h}-\boldsymbol{y}\right\|_2^2\right\} \quad (3.9)$$

where the two separate 4-parameter PAR models are integrated into one least-squares problem with weight $\xi^{\times}$ and ξ^{+} namely,

$$\xi^{\times}=\frac{e^{+}}{e^{+}+e^{\times}} \qquad \xi^{+}=\frac{e^{\times}}{e^{+}+e^{\times}} \quad (3.10)$$

The two weights $\xi^{\times}$ and ξ^{+} are derived from squared error e^{+} and $e^{\times}$ which are associated with corresponding PAR models in solving the least-squares problems in Eq. 3.2. They account for the fitting degree of the PAR models to pixels in the local window W. Term λ is the Lagrange multiplier, and term $\lambda\left\|\boldsymbol{x}*\boldsymbol{h}-\boldsymbol{y}\right\|_2^2$ imposes a constraint on the solution space of the ill-posed problem. The operator $*$ denotes the cascaded operations of low-pass filtering and down-sampling. It corresponds to the physical formation of observed low-resolution pixels $\boldsymbol{y}$ from the underlying high-resolution pixels $\boldsymbol{x}$ in local window W. Term $\boldsymbol{h}$ accounts for the PSF effects of the digital imaging system that plays a role of low-pass filtering. The PSF function can be estimated from the imaging system[1]. In this thesis, it is assumed to be known as a 3 × 3 Gaussian kernel:

$$p(x,y)=\begin{bmatrix}0 & 1 & 0\\ 1 & 4 & 1\\ 0 & 1 & 0\end{bmatrix} \quad (3.11)$$

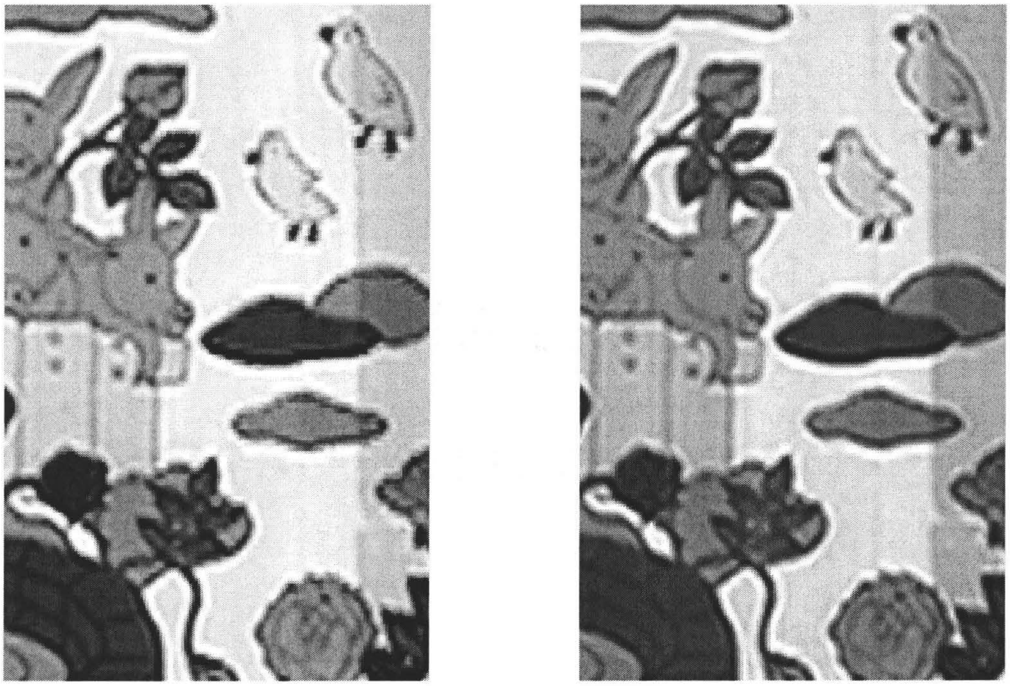
3.5 Simulation Results and Discussion

In this section, we evaluate the performance of our proposed super-resolution method in comparison with one recently published super-resolution counterpart[37] on natural video sequences, namely *calendar*, *car* and *foreman*. The testing video sequences include a variety of motion types such as translation, rotation and scaling etc. In *calendar* and *foreman* sequence, there exist multiple local motions in the scene. Image resolution of all the video frames is upconverted by a factor of two. The comparison in terms of visual quality is shown by Fig. 3.5, 3.6 and 3.7. Compared with the results by [37], high-resolution video frames reconstructed by our method have fewer jaggies on the edge features. Fig. 3.8 shows the comparison of our proposed method with the method in [37] in terms of PSNR measurement on *foreman* sequence. The testing video sequence is prefilterred by a low-pass filter as shown in Eq. 3.11 prior to uniform down-sampling. After that, image resolution of the down-sampled video frames is upconverted by a factor of two, using the proposed method and the method in [37] respectively. Based on the comparison in Fig. 3.8, we can conclude that the proposed method surpasses the method in [37] by 1-2dB over 65 of the 80 testing video frames.

3.6 Conclusion

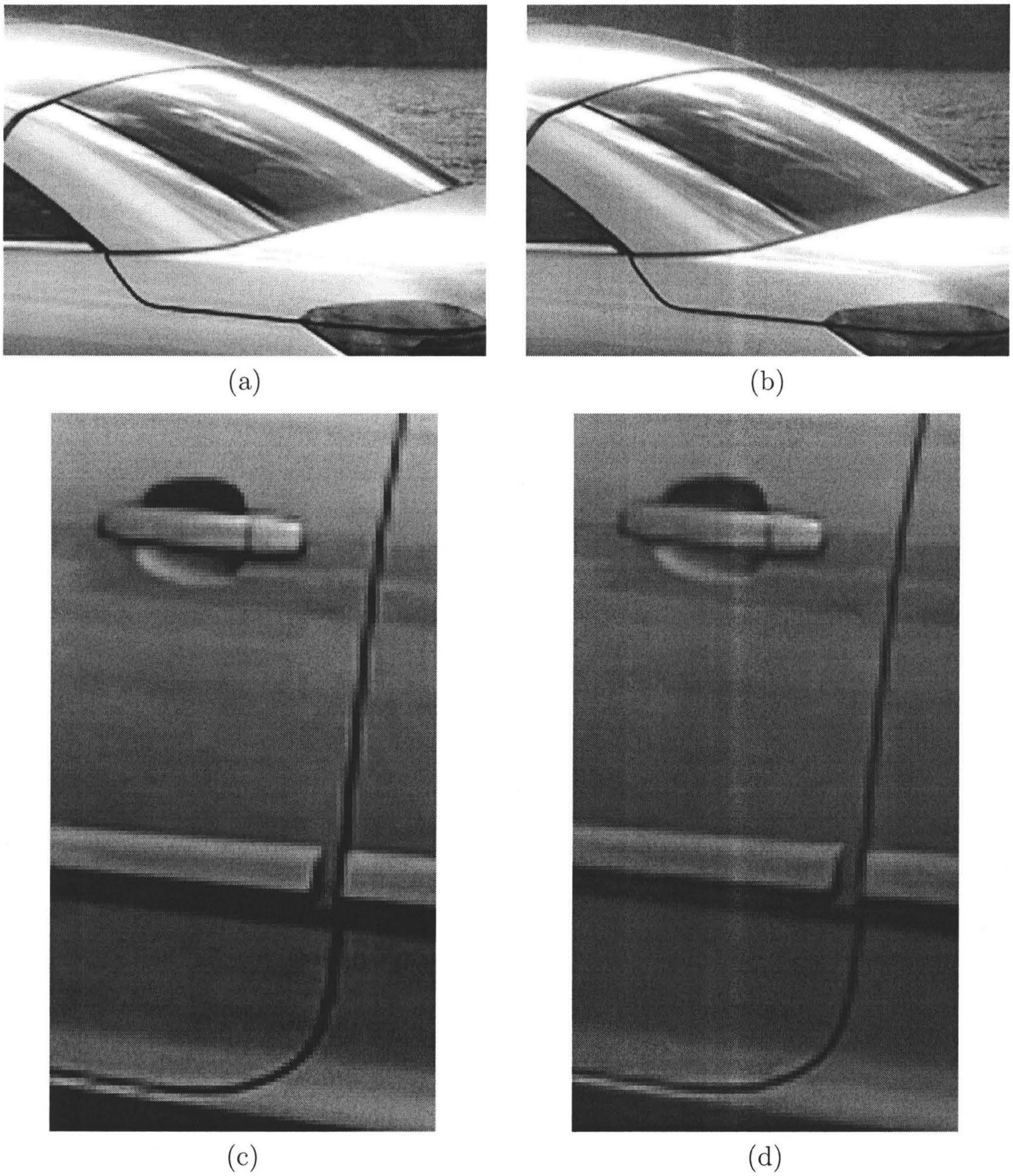
In this chapter, a new model-learning scheme is proposed such that PAR models can be learned from observed data of multiple registered low-resolution video frames. The proposed scheme takes advantage of the temporal correlation of video frames. It integrates pixel samples from multiple registered low-resolution frames into a local window. Therefore, it increases pixel density of the local window without violating the piecewise stationary nature of image signals. Then by performing model learning in the high-resolution local window, the new scheme improves the accuracy of PAR model parameters. Further, it reduces the possibility of PAR model mismatch between the observed low-resolution frame and the underly high-resolution frame. VSR reconstruction in this chapter is formulated as a model-based still image interpolation problem. By using the estimated PAR model parameters, missing pixels of the underlying high-resolution video frames are reconstructed by solving a linear least-squares problem.

In addition, our proposed method utilizes the most common non-parametric Block-Matching Algorithm for motion estimation. Compared with conventional counterparts that employ parametric motion models, our VSR method reduces the computing complexity of motion estimation.



(a) (b)

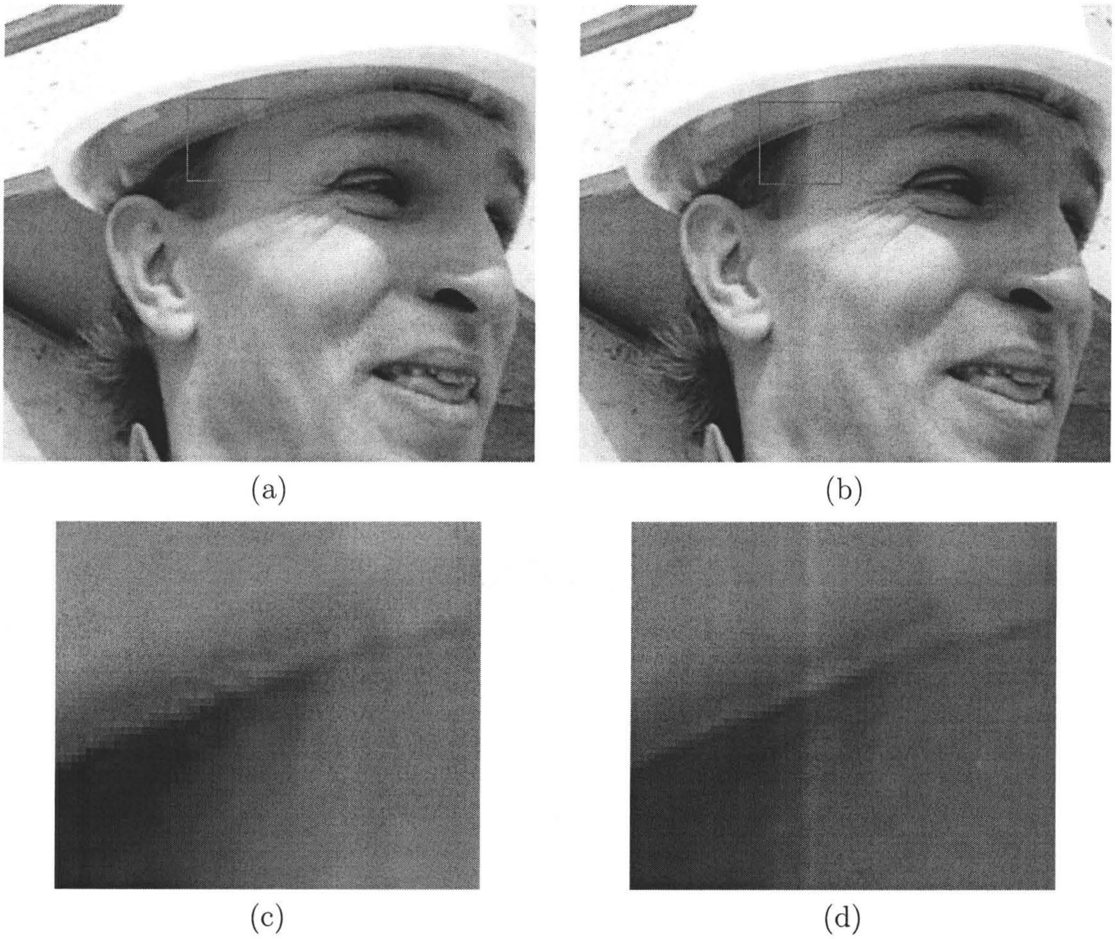
Figure 3.5: Comparison on Calendar sequence: (a) result by [37]; (b) result by proposed method.



(a) (b)

(c) (d)

Figure 3.6: Comparison on Car sequence: (a)(c) result by [37]; (b)(d) result by proposed method.



(a) (b)

(c) (d)

Figure 3.7: Comparison on Foreman sequence: (a)(c) results by [37]; (b)(d) results by proposed method.

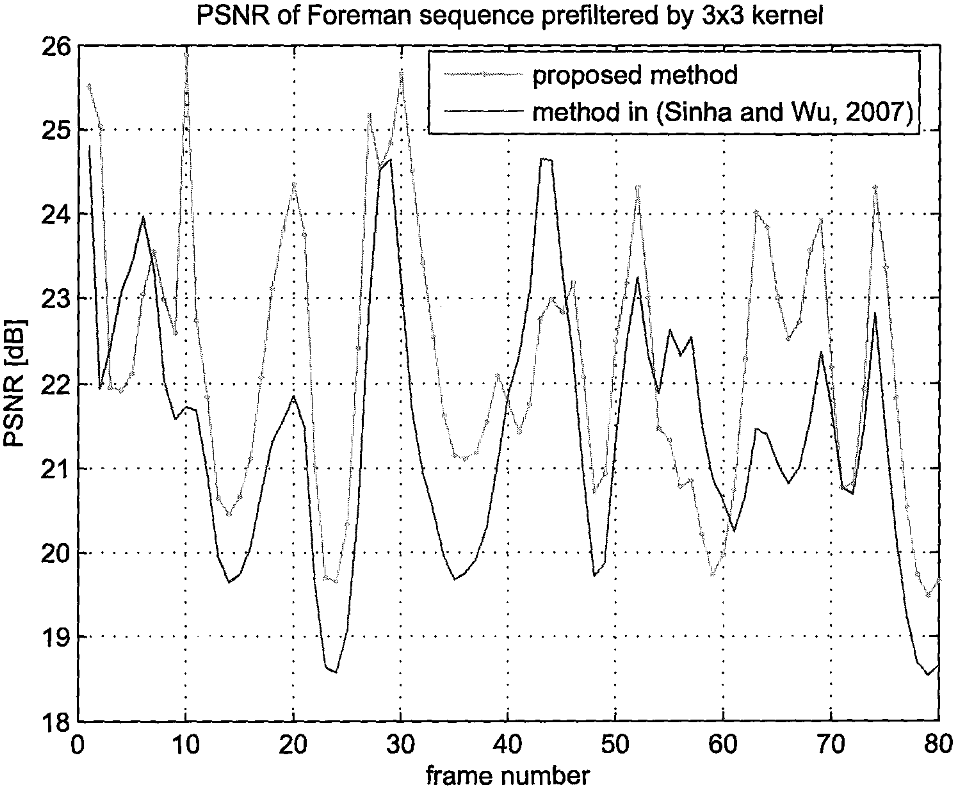


Figure 3.8: Comparison of proposed method with the counterpart in [37] on *foreman* sequence.

Chapter 4

PAR Model based Joint Motion Estimation and Video Super-Resolution

In the preceding chapter, we dealt with the VSR problem in an approach of model-based image interpolation. This approach does not reconstruct the high-resolution frames by solving the expensive inverse problem of VSR that involves motion-registered multiple frames. The data redundancy afforded by motions is only used to estimate the parameters of the PAR model, and improve the accuracy and robustness of the PAR model. The resulting PAR model is used to interpolate the low-resolution frame in the spatial domain. This simplified interpolation-type VSR technique has the advantage of low computational complexity, but its performance is not as good as the classical linear algebraical approach as outlined in Section 2.2.

In this chapter, we reexamine the well-known VSR inverse problem derived from the multi-frame observation model, and propose an improved PAR model based solution. As ananlyzed in Chapter 2, observation model based VSR reconstruction is an ill-posed inverse problem whose solution heavily depends on the choice of a regularization term. Historically, most existing VSR algorithms in the literature use total variation (TV) for the regularization term[9]. As a matter of fact, TV-based image restoration methods can not effectively preserve subtle image details. As such VSR algorithms that use TV for VSR regularization are incapable of reproducing high-frequency components of image signals either.

Motivated by the effectiveness of a PAR model based still image interpolation technique[44], we propose to apply the PAR model to replacing the most common regularization term of TV in regularizing solutions for the VSR problem. It makes the estimation of the missing HR pixel intensities as minimizing the following Lagrangian.

$$\hat{\boldsymbol{z}} = \arg \min_{\boldsymbol{z}} \left\{ \left\| \boldsymbol{g} - \boldsymbol{F}\boldsymbol{z} \right\|_2^2 + \lambda \left\| \boldsymbol{A}\boldsymbol{z} \right\|_2^2 \right\} \tag{4.1}$$

where vectors $\boldsymbol{z}$ and $\boldsymbol{g}$ denote the underlying HR pixel intensities and observed LR pixel intensities in lexicographical order respectively. Matrix $\boldsymbol{F}$ represents joint operations of geometry warping, blurring and down-sampling as discussed in Chapter 2. Term $\lambda \left\| \boldsymbol{A}\boldsymbol{z} \right\|_2^2$ imposes a constraint on the solution, suggesting the spatial coherence of pixel structures between the underlying HR image and observed LR images. Term λ is the Lagrange multiplier that provides a balance between the fidelity to the underlying data (as expressed by $\left\| \boldsymbol{g} - \boldsymbol{F}\boldsymbol{z} \right\|_2^2$) and coherence of the solution with observed data (as expressed by $\lambda \left\| \boldsymbol{A}\boldsymbol{z} \right\|_2^2$). Term $\boldsymbol{A}$ serves as a linear PAR model parameter matrix. Each row of vector $\boldsymbol{A}\boldsymbol{z}$ is a form of $\sum_k \{a_k z_{i,k} - z_i\}$, where a_k is a PAR model parameter associated with one pixel $z_{i,k}$ in the neighborhood of pixel z_i.

Compared with TV-based methods, the PAR model based regularization method has a distinctive advantage of adaptivity to local image statistics. It can, by spatially varying its parameters, adapt the reconstruction of high-resolution frames to local image waveforms. Furthermore, it can effectively preserve the spatial coherence of image structures by fitting the underlying high-resolution pixels to the PAR model that is learnt from the observed low-resolution image. However, as analyzed in Chapter 3, inconsistent second-order statistics of low-resolution image and high-resolution image can lead to a pitfall of PAR model mismatch. The model mismatch potentially downgrades the performance of the PAR model in image resolution upconversion.

Additionally, in this addressed VSR reconstruction problem, there are three groups of unknown variables namely, motion parameters, missing high-resolution pixel intensities and the parameters of the PAR model. The reconstruction of high-resolution frames depends on the estimated PAR model parameters and motion parameters, and vice versa. It suggests that computing these unknowns is a problem with chicken-and-egg flavor. In view of this, we propose a new VSR method that estimates the PAR model parameters, the motion parameters and the underlying high-resolution frames jointly. Our new method can improve the accuracy of PAR model parameters by learning them from reconstructed HR frames iteratively. Furthermore, it can mitigate motion estimation errors that exist in computing subpixel precision motion parameters. Even though some previous papers[13, 14] proposed similar ideas in joint estimation, they coincidentally use TV methods for regularization and hence can not avoid the significant loss of fine image details. At the end of this chapter, simulation results show that our VSR method gains competitive performance in terms of visual quality.

The initial values of unknown variables are critical to the performance of an iterative VSR algorithm. In our method, we estimate the initial values of motion parameters and PAR model parameters from LR frames due to the absence of HR version. Furthermore, we propose to compute increments for all the unknowns and make them simultaneously approach the optimal solutions. Therefore, we can achieve best statistical consistency among the three groups of unknown variables.

The rest of this paper is organized as follows. In Section 4.1, we formulate the VSR reconstruction problem. In Section 4.2 and 4.3, we derive two Jacobean matrices. In Section 4.4, we present an iterative solution for the nonlinear least-squares problem. In Section 4.5, we present simulation results on nature video sequences.

4.1 Problem Formulation

Since there are three groups of unknown variables in the addressed VSR problem, the Lagrangian in Eq. 4.1 can be rewritten into the following nonlinear least-squares (LS) form.

$$\min_{\{\boldsymbol{\alpha},\boldsymbol{\beta}\},\boldsymbol{\nu},\boldsymbol{z}} \left\{(\boldsymbol{g}-\boldsymbol{F}(\boldsymbol{\nu})\cdot\boldsymbol{z})^T(\boldsymbol{g}-\boldsymbol{F}(\boldsymbol{\nu})\cdot\boldsymbol{z})+\lambda(\boldsymbol{A}(\boldsymbol{\alpha},\boldsymbol{\beta})\cdot\boldsymbol{z})^T(\boldsymbol{A}(\boldsymbol{\alpha},\boldsymbol{\beta})\cdot\boldsymbol{z})\right\} \tag{4.2}$$

where vector $\boldsymbol{\nu}$ represents motion parameters as described in Section 2.1. Matrix $\boldsymbol{A}$ is derived from recomposing the autoregressive model expression of Eq. 3.1 into a matrix-vector form. Therefore, it is composed of PAR model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. For the simplicity of notation, hereafter we have vector $\boldsymbol{a}=(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)^T$, and thus

$\boldsymbol{A}(\boldsymbol{a}) = \boldsymbol{A}(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Let us define a cost function:

$$\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{a}, \boldsymbol{\nu}, \boldsymbol{z}) = \begin{pmatrix} \boldsymbol{g} \\ \boldsymbol{0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{F}(\boldsymbol{\nu}) \\ \sqrt{\lambda}\boldsymbol{A}(\boldsymbol{a}) \end{pmatrix} \cdot \boldsymbol{z} \tag{4.3}$$

where term $\boldsymbol{0}$ is a zero vector. Then the nonlinear LS problem in Eq.4.2 can be reformulated as

$$\min_{\boldsymbol{a},\boldsymbol{\nu},\boldsymbol{z}} \boldsymbol{r}^T \boldsymbol{r} \tag{4.4}$$

In what follows, we will derive an iterative scheme that simplifies the minimization of the nonlinear LS as iterative minimization of a linear LS. First of all, let us take into account increments $\triangle\boldsymbol{a}$, $\triangle\boldsymbol{\nu}$ and $\triangle\boldsymbol{z}$ with respect to the corresponding vector $\boldsymbol{a}$, $\boldsymbol{\nu}$ and $\boldsymbol{z}$. Then cost function $\boldsymbol{r}$ is updated as $\boldsymbol{r}(\boldsymbol{a} + \triangle\boldsymbol{a}, \boldsymbol{\nu} + \triangle\boldsymbol{\nu}, \boldsymbol{z} + \triangle\boldsymbol{z})$. Based on Eq. 4.3, we can get a linear form of Eq. 4.5 and 4.6 by expanding $\boldsymbol{r}(\boldsymbol{a} + \triangle\boldsymbol{a}, \boldsymbol{\nu} + \triangle\boldsymbol{\nu}, \boldsymbol{z} + \triangle\boldsymbol{z})$ via Taylor's expansion and ignoring the second and higher order terms.

$$\begin{aligned} & \boldsymbol{g} - \boldsymbol{F}(\boldsymbol{\nu} + \triangle\boldsymbol{\nu}) \cdot (\boldsymbol{z} + \triangle\boldsymbol{z}) \\ = & \ \boldsymbol{g} - \boldsymbol{F}(\boldsymbol{\nu} + \triangle\boldsymbol{\nu}) \cdot \boldsymbol{z} - \boldsymbol{F}(\boldsymbol{\nu} + \triangle\boldsymbol{\nu}) \cdot \triangle\boldsymbol{z} \\ \approx & \ \boldsymbol{g} - \boldsymbol{F}(\boldsymbol{\nu}) \cdot \boldsymbol{z} - \boldsymbol{J}_\nu \cdot \triangle\boldsymbol{\nu} - \boldsymbol{F}(\boldsymbol{\nu}) \cdot \triangle\boldsymbol{z} \end{aligned} \tag{4.5}$$

$$\begin{aligned} & \boldsymbol{A}(\boldsymbol{a} + \triangle\boldsymbol{a}) \cdot (\boldsymbol{z} + \triangle\boldsymbol{z}) \\ = & \ \boldsymbol{A}(\boldsymbol{a} + \triangle\boldsymbol{a}) \cdot \boldsymbol{z} + \boldsymbol{A}(\boldsymbol{a} + \triangle\boldsymbol{a}) \cdot \triangle\boldsymbol{z} \\ \approx & \ \boldsymbol{A}(\boldsymbol{a}) \cdot \boldsymbol{z} + \boldsymbol{J}_a \cdot \triangle\boldsymbol{a} + \boldsymbol{A}(\boldsymbol{a}) \cdot \triangle\boldsymbol{z} \end{aligned} \tag{4.6}$$

where $\boldsymbol{J}_\nu$, $\boldsymbol{J}_a$ are the Jacobian matrix of $\boldsymbol{F}(\boldsymbol{\nu})\boldsymbol{z}$ and $\boldsymbol{A}(\boldsymbol{a})\boldsymbol{z}$ respectively.

Therefore, cost function $\boldsymbol{r}$ with minor change with respect to $\boldsymbol{a}$, $\boldsymbol{\nu}$ and $\boldsymbol{z}$ becomes a linear function of $\triangle\boldsymbol{a}$, $\triangle\boldsymbol{\nu}$ and $\triangle\boldsymbol{z}$ namely,

$$\begin{aligned} & \boldsymbol{r}(\boldsymbol{a}+\triangle\boldsymbol{a},\boldsymbol{\nu}+\triangle\boldsymbol{\nu},\boldsymbol{z}+\triangle\boldsymbol{z}) \\ = \; & \boldsymbol{r}(\boldsymbol{a},\boldsymbol{\nu},\boldsymbol{z}) - \begin{pmatrix} 0 & \boldsymbol{J}_v & \boldsymbol{F}(\boldsymbol{\nu}) \\ \sqrt{\lambda}\boldsymbol{J}_a & 0 & \sqrt{\lambda}\boldsymbol{A}(\boldsymbol{a}) \end{pmatrix} \begin{pmatrix} \triangle\boldsymbol{a} \\ \triangle\boldsymbol{\nu} \\ \triangle\boldsymbol{z} \end{pmatrix} \end{aligned} \tag{4.7}$$

Then computing increments $\triangle\boldsymbol{a}$, $\triangle\boldsymbol{\nu}$ and $\triangle\boldsymbol{z}$ is formulated as minimizing L_2 norm of cost function $\boldsymbol{r}(\boldsymbol{a}+\triangle\boldsymbol{a},\boldsymbol{\nu}+\triangle\boldsymbol{\nu},\boldsymbol{z}+\triangle\boldsymbol{z})$ given $\boldsymbol{r}(\boldsymbol{a},\boldsymbol{\nu},\boldsymbol{z})$. In the proceeding two sections, we will derive Jacobian matrix $\boldsymbol{J}_\nu$ and $\boldsymbol{J}_a$ such that these increments can be computed by solving a linear LS problem.

4.2 Derivation of Jacobean Matrix of $\boldsymbol{F}(\boldsymbol{\nu})\boldsymbol{z}$

In this chapter, we focus on a special case where the block-based reconstruction is employed and geometry warping is pure translational motion. Specifically, motion parameters are identical within each block. Define $(\boldsymbol{x}_0, \boldsymbol{y}_0)$, $(\boldsymbol{x}_k, \boldsymbol{y}_k)$ as pixel coordinates of the current frame and k-th reference frame in lexicographical order respectively, where $k = 1, 2, \cdots, N$. Then the displacement between the two blocks of pixels is $\boldsymbol{\nu}_k = (d\boldsymbol{x}_k, d\boldsymbol{y}_k)$, where

$$d\boldsymbol{x}_k = \boldsymbol{x}_k - \boldsymbol{x}_0, \quad d\boldsymbol{y}_k = \boldsymbol{y}_k - \boldsymbol{y}_0 \tag{4.8}$$

Accordingly, we can easily get

$$\boldsymbol{J}_\nu = \frac{\partial(\boldsymbol{F}(\boldsymbol{\nu})\boldsymbol{z})}{\partial\boldsymbol{\nu}} = \frac{\partial(\boldsymbol{F}(\boldsymbol{\nu})\boldsymbol{z})}{\partial([\boldsymbol{dx},\boldsymbol{dy}])} = \left(\frac{\partial(\boldsymbol{F}(\boldsymbol{\nu})\boldsymbol{z})}{\partial(\boldsymbol{dx})}, \frac{\partial(\boldsymbol{F}(\boldsymbol{\nu})\boldsymbol{z})}{\partial(\boldsymbol{dy})}\right) \tag{4.9}$$

In the case of image resolution upconversion by a factor of two, we use an area-based interpolation technique based on the assumption of box PSF[37]. As shown in Fig. 4.1, the LR pixel represented by a dashed rectangle is registered onto an HR grid where the underlying HR pixels are denoted by smaller solid rectangles. Intensity value of the LR pixel is predicted by a linear combination of the underlying HR pixel $\boldsymbol{P}$ with coefficient $\boldsymbol{C}$. $\boldsymbol{C}$ is a vector and each of its elements is proportional to the area of overlap with the associated HR pixel $\boldsymbol{P}$, where

$$\begin{aligned}\boldsymbol{C} =& ((1-dx)(1-dy), 1-dx, (1-dy)dy,\\ & 1-dy, 1, dy, dx(1-dy), dx, dxdy)\\ \boldsymbol{P} =& (I_0, I_1, I_2, I_3, I_4, I_5, I_6, I_7, I_8)\end{aligned} \tag{4.10}$$

Hence, each element of vector $\boldsymbol{F}(\boldsymbol{\nu})\boldsymbol{z}$, corresponding to one observed LR pixel, equals

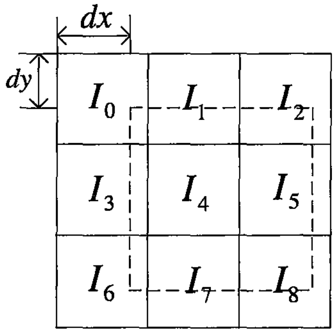


Figure 4.1: Illustration of a LR pixel projected onto an HR grid.

the inner product of $\boldsymbol{C}$ and $\boldsymbol{P}$ which is denoted as $< \boldsymbol{C}, \boldsymbol{P} >$. Most importantly, it is easy to find that $\boldsymbol{F}(\boldsymbol{\nu})\boldsymbol{z}$ is linear and differentiable with respect to motion parameter $\boldsymbol{\nu}$.

Since the observed LR pixels are independent from each other, we can formulate Jacobian matrix $\boldsymbol{J}_\nu$ as

$$\boldsymbol{J}_\nu = \begin{pmatrix} \boldsymbol{J}_{x,0} & \cdots & 0 & \boldsymbol{J}_{y,0} & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{J}_{x,N-1} & 0 & \cdots & \boldsymbol{J}_{y,N-1} \end{pmatrix} \tag{4.11}$$

where $\boldsymbol{J}_{x,s}$ and $\boldsymbol{J}_{y,s}$ are the Jacobian matrix of $\boldsymbol{F}(s, \boldsymbol{\nu}_s)\boldsymbol{z}$, $s = 0, 1, \cdots, N-1$. N is the number of observed LR frames. $\boldsymbol{J}_{x,s}$ and $\boldsymbol{J}_{y,s}$ are derived as follows.

$$\begin{aligned} \boldsymbol{J}_{x,s} &= \frac{\partial(<\boldsymbol{C}, \boldsymbol{P}>)}{\partial(dx)} = (1-dy)(I_6 - I_0) + dy(I_8 - I_2) + I_7 - I_1 \\ \boldsymbol{J}_{y,s} &= \frac{\partial(<\boldsymbol{C}, \boldsymbol{P}>)}{\partial(dy)} = (1-dx)(I_6 - I_0) + dx(I_8 - I_2) + I_7 - I_1 \end{aligned} \tag{4.12}$$

Note that vector $\boldsymbol{P}$ is determined by the underlying HR pixels, and therefore its vector value differs in pixels of the HR grid.

4.3 Derivation of Jacobean Matrix of $\boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{z}$

The Jacobian matrix of $\boldsymbol{A}(\boldsymbol{a})\boldsymbol{z}$, is

$$\boldsymbol{J}_a = \frac{\partial(\boldsymbol{A}(\boldsymbol{a})\boldsymbol{z})}{\partial([\boldsymbol{\alpha}, \boldsymbol{\beta}])} = \begin{pmatrix} \frac{\partial(\boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{z})}{\partial\boldsymbol{\alpha}} & \boldsymbol{0} \\ \boldsymbol{0} & \frac{\partial(\boldsymbol{A}(\boldsymbol{\beta})\boldsymbol{z})}{\partial\boldsymbol{\beta}} \end{pmatrix} \tag{4.13}$$

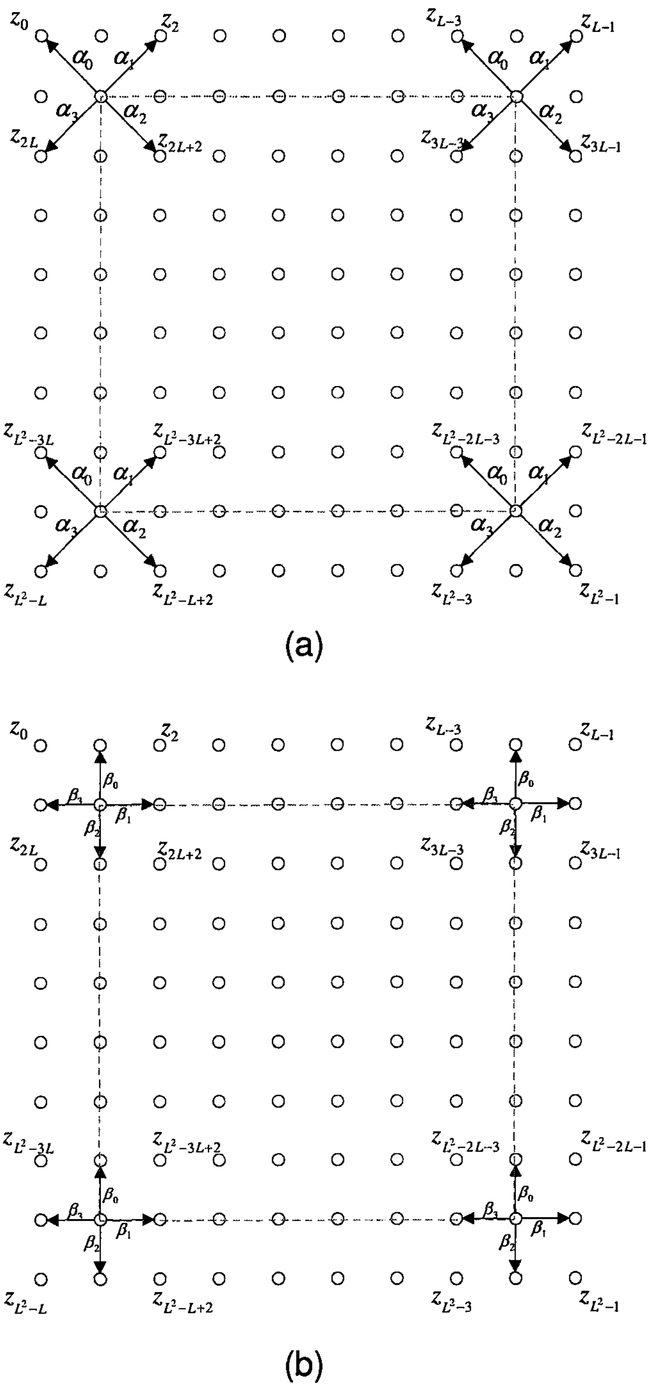


Figure 4.2: Illustration of vector $\boldsymbol{Az}$ in a local window of size $L \times L$: (a) diagonal mode, (b) axial mode. Each pixel within the dashed rectangles in (a) and (b) corresponds to an element in vector $\boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{z}$ and $\boldsymbol{A}(\boldsymbol{\beta})\boldsymbol{z}$ respectively.

where $\boldsymbol{A}(\boldsymbol{a}) = [\boldsymbol{A}^T(\boldsymbol{\alpha}), \boldsymbol{A}^T(\boldsymbol{\beta})]^T$. Each element of vector $\boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{z}$ and $\boldsymbol{A}(\boldsymbol{\beta})\boldsymbol{z}$ equals $\sum\limits_{0\leq k\leq 3}\{\alpha_k z^{\times}_{i,k} - z_i\}$ and $\sum\limits_{0\leq k\leq 3}\{\beta_k z^{+}_{i,k} - z_i\}$ respectively. Fig. 4.2 illustrates vector $\boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{z}$ and $\boldsymbol{A}(\boldsymbol{\beta})\boldsymbol{z}$ in a local window of size $L \times L$. $\boldsymbol{A}(\boldsymbol{a})\boldsymbol{z}$ is linear and differentiable with respect to PAR model parameter $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Accordingly, Jacobian matrix $\boldsymbol{J}_a$ can be represented in the form of the underlying high-resolution pixel intensity $\boldsymbol{z}$ namely,

$$\frac{\partial(\boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{z})}{\partial\boldsymbol{\alpha}} = \begin{pmatrix} z_0 & z_2 & z_{2L+2} & z_{2L} \\ z_1 & z_3 & z_{2L+3} & z_{2L+1} \\ \vdots & \vdots & \vdots & \vdots \\ z_{L^2-2L-3} & z_{L^2-2L-1} & z_{L^2-1} & z_{L^2-3} \end{pmatrix} \tag{4.14}$$

$$\frac{\partial(\boldsymbol{A}(\boldsymbol{\beta})\boldsymbol{z})}{\partial\boldsymbol{\beta}} = \begin{pmatrix} z_1 & z_{L+2} & z_{2L+1} & z_L \\ z_2 & z_{L+3} & z_{2L+2} & z_{L+1} \\ \vdots & \vdots & \vdots & \vdots \\ z_{L^2-2L-2} & z_{L^2-L-1} & z_{L^2-2} & z_{L^2-L-3} \end{pmatrix} \tag{4.15}$$

4.4 Iterative Solution for the Nonlinear Total Least-Square Problem

Given the derivation of Jacobian matrix $\boldsymbol{J}_a$ and $\boldsymbol{J}_v$ above as well as initial estimates of vectors $\boldsymbol{a}$, $\boldsymbol{\nu}$ and $\boldsymbol{z}$, the increments $\triangle\boldsymbol{a}$, $\triangle\boldsymbol{\nu}$ and $\triangle\boldsymbol{z}$ can be obtained by solving the following linear LS problem.

$$\min_{\triangle \boldsymbol{a}, \triangle \boldsymbol{\nu}, \triangle \boldsymbol{z}} \left\| \begin{pmatrix} \boldsymbol{r}(\boldsymbol{a}, \boldsymbol{\nu}, \boldsymbol{z}) \\ 0 \end{pmatrix} - \begin{pmatrix} 0 & \boldsymbol{J}_v & \boldsymbol{F}(\boldsymbol{\nu}) \\ \sqrt{\lambda}\boldsymbol{J}_a & 0 & \sqrt{\lambda}\boldsymbol{A}(\boldsymbol{a}) \\ \boldsymbol{R}_1 & 0 & 0 \\ 0 & \boldsymbol{R}_2 & 0 \end{pmatrix} \begin{pmatrix} \triangle \boldsymbol{a} \\ \triangle \boldsymbol{\nu} \\ \triangle \boldsymbol{z} \end{pmatrix} \right\|_2^2 \tag{4.16}$$

where $\boldsymbol{R}_1 = \lambda_1 \boldsymbol{I}$ and $\boldsymbol{R}_2 = \lambda_2 \boldsymbol{I}$ are two regularization matrices that stabilize the linear LS problem in Eq. 4.16. Term λ_1, λ_2 are regularization constants, and $\boldsymbol{I}$ is an identity matrix whose size is adjusted in accordance with vectors $\boldsymbol{a}$ and $\boldsymbol{\nu}$.

In the end, we summarize our block-based iterative method for joint estimation of the increments $\triangle \boldsymbol{a}$, $\triangle \boldsymbol{\nu}$ and $\triangle \boldsymbol{z}$ in the following Algorithm 1. It should be pointed out that the initial values of PAR model parameter $\boldsymbol{a}$ are estimated from LR frames by solving the linear LS problem as shown in Eq. 3.2. Subpixel level motion parameters are initialized by performing block-matching algorithm (BMA) between two neighboring low-resolution video frames. The algorithms converges when the average of pixel intensity increment $\triangle \boldsymbol{z}$ is bounded within one pixel.

4.5 Simulation Results and Discussion

The results of our proposed super-resolution method are illustrated in Fig. 4.3-4.5 in comparison with one recently published counterpart[37] on natural video sequences. The VSR method in [37] employs the same observation image model as our proposed method. But, in contrast, it uses bicubic interpolated HR results as constraints to regularize the VSR solution[37]. Moreover, it employs the conventional two-step VSR

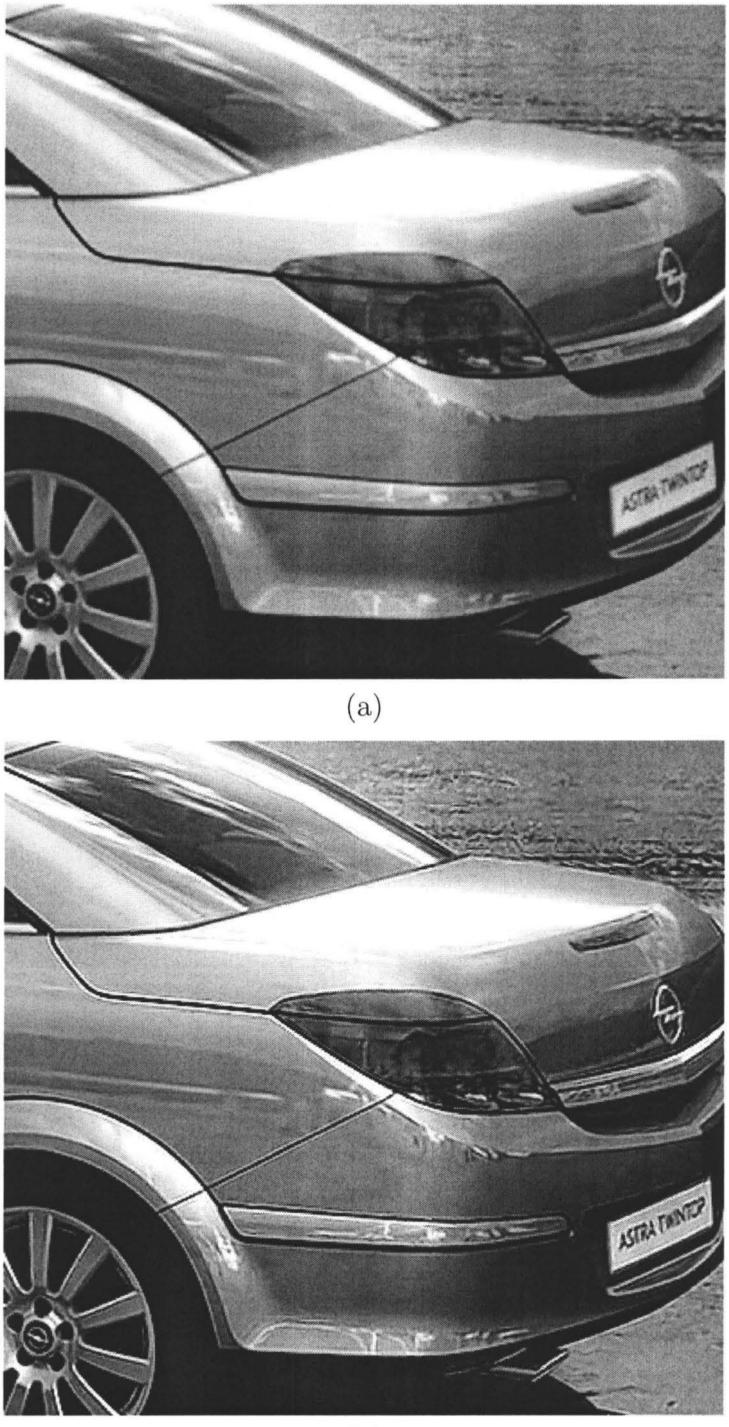
Algorithm 1 Proposed Iterative Algorithm

```
1: For each block of a frame, initially estimate vector $\boldsymbol{\nu}^{(0)}$, $\boldsymbol{a}^{(0)}$ from LR frames using
   above-mentioned methods, and then compute the underlying HR pixel intensities
   $\boldsymbol{z}^{(0)}$ by Eq. 4.1, as well as the cost function $\boldsymbol{r}^{(0)}$ in Eq. 4.3.
2: Compute Jacobian matrix $\boldsymbol{J}_\nu$ and $\boldsymbol{J}_a$ by Eq. 4.11 and 4.13 prior to computing
   increments $\triangle\boldsymbol{a}^{(i)}$, $\triangle\boldsymbol{\nu}^{(i)}$ and $\triangle\boldsymbol{z}^{(i)}$ by Eq. 4.16.
3: Update terms:
       $\boldsymbol{\nu}^{(i+1)} = \boldsymbol{\nu}^{(i)} + \triangle\boldsymbol{\nu}^{(i)},$
       $\boldsymbol{a}^{(i+1)} = \boldsymbol{a}^{(i)} + \triangle\boldsymbol{a}^{(i)},$
       $\boldsymbol{z}^{(i+1)} = \boldsymbol{z}^{(i)} + \triangle\boldsymbol{z}^{(i)}$
   and $\boldsymbol{r}(\boldsymbol{a}^{(i+1)}, \boldsymbol{\nu}^{(i+1)}, \boldsymbol{z}^{(i+1)})$ according to Eq. 4.3, 4.5 and 4.6.
4: Go to Step 2 repeatedly until $\triangle\boldsymbol{a}^{(i)}$, $\triangle\boldsymbol{\nu}^{(i)}$ $\triangle\boldsymbol{z}^{(i)}$ and $\boldsymbol{r}(\boldsymbol{a}^{(i)}, \boldsymbol{\nu}^{(i)}, \boldsymbol{z}^{(i)})$ satisfy the
   converging requirement, or it reaches a maximum number of iterations.
5: Achieve an optimal solution for underlying HR pixel $\boldsymbol{z}$ in one block.
6: Go to Step 1 until the whole HR frame is reconstructed.
```

scheme (as mentioned in Section 2.2), and uses a 6-parameter affine motion model. The testing video frames shown in the figures are part of the original frames which are upconverted by a factor of two. In *Car* sequence, there exist significant translation motions and slight zooming motions. In *Foreman* sequence, the wall in the first half of the sequence and the tile in the rear part of the sequence, have significant translation motions as well as slight rotation motions. Based on the simulation results, we can conclude that our proposed method gains superior performance in preserving high-frequency components of images.

4.6 Conclusion

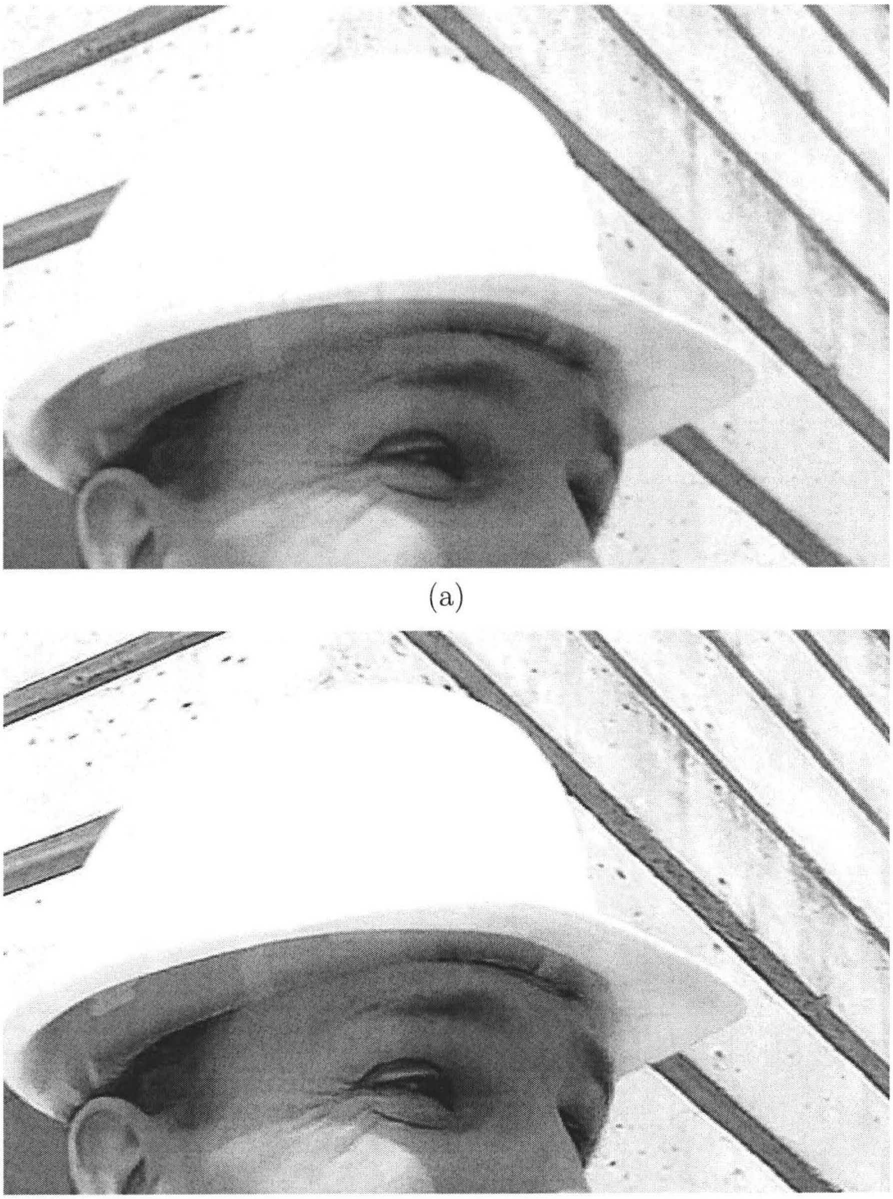
In this chapter, a joint super-resolution method based on a piecewise autoregressive image model has been proposed for video super-resolution reconstruction. The contribution of the proposed method lies in two aspects. First, the autoregressive image model is introduced as a regularization term to replace the commonly used total variation method for super-resolution reconstruction. The proposed regularization term, by spatially adapting to local image waveforms, effectively preserves the pixel structure of image signals. Secondly, an iterative scheme for joint estimation of motion parameters, autoregressive model parameters and missing high-resolution pixels are proposed. The joint method significantly mitigates estimation errors and the possibility of model mismatch. The effectiveness of the proposed method in super-resolution has been demonstrated by simulation results on natural video sequences.



(a)

(b)

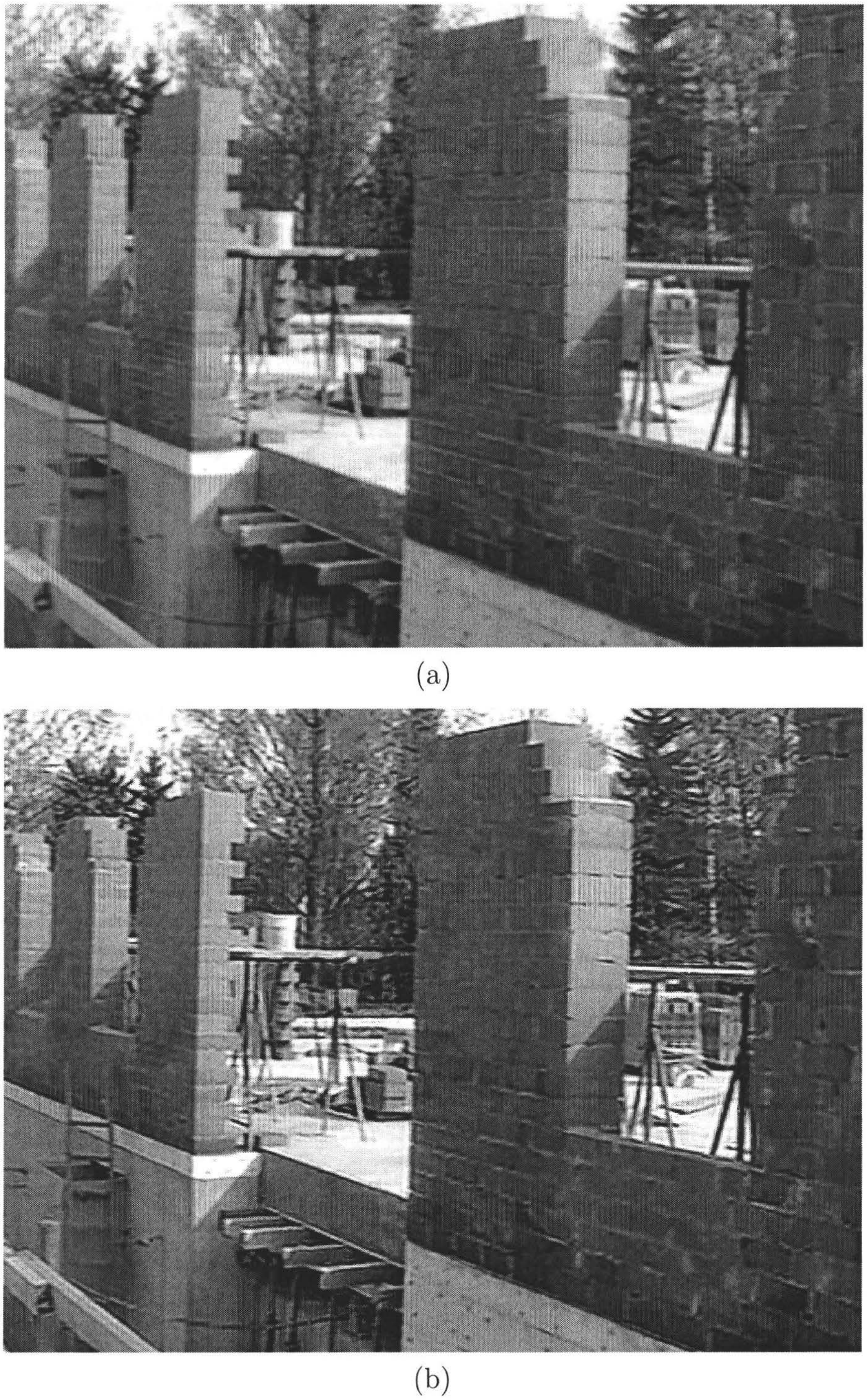
Figure 4.3: Comparison on Car sequence: (a) result by [37]; (b) result by proposed method.



(a)

(b)

Figure 4.4: Comparison on Foreman sequence: (a) result by [37]; (b) result by proposed method.



(a)

(b)

Figure 4.5: Comparison on Foreman sequence: (a) result by [37]; (b) result by proposed method.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this thesis, we investigated the classical problem of video super-resolution and proposed two new methods which were able to effectively improve visual quality of reconstructed video frames. Both the two proposed methods take the advantage of a piecewise 2D autoregressive image model, but address the super-resolution problem in distinct manners. In the first method, we propose a new model-learning scheme that learns the model parameters from pixel samples of multiple registered low-resolution video frames. This scheme not only reduces the possibility of model mismatch between the low-resolution frame and high-resolution frame, but also increases the accuracy of estimated model parameters, which hence improves the performance of the autoregressive model in high-resolution pixel reconstruction. In the second method, we formulate super-resolution reconstruction problem via an observation model, and incorporate the autoregressive model to regularize solutions for the ill-posed inverse problem. This method can preserve the spatial coherence of pixel structures in the reconstructed high-resolution video frames. Furthermore, both the proposed methods are adept at preserving image details by adaptively varying the reconstruction of high-resolution pixels to local image waveforms. Simulations have been conducted and the results convincingly demonstrate the competitive performance of the proposed methods in comparison with its counterparts.

5.2 Future Work

Even though our methods achieved competitive results in video super-resolution, there are still one issue remaining to be addressed in the future. As discussed in [44], the estimation of autoregressive model parameters is a challenging task. In this thesis, our methods improve the estimation accuracy of model parameters by registering pixels of multiple low-resolution video frames onto a high-resolution grid to increase the pixel density. However, there is a possibility that some of the pixel sites on the high-resolution grid can be not filled. It results in inconsistent pixel distances in various directions of the autoregressive model. At present, we circumvent this problem by ignoring its negative effects. But, a more valid method should provide a unified framework in estimating the autoregressive model parameters. In the future, any progress in this regards would further strengthen the effectiveness of the autoregressive model.

Bibliography

[1] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, 2002.

[2] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.

[3] S. Borman and R. L. Stevenson. Super-resolution from image sequences-a review. In *Proc. Midwest Symposium on Circuits and Systems*, pages 374–378, 9–12 Aug. 1998.

[4] M. Born, E. Wolf, and A. Bhatia. *Principles of optics.* Pergamon press New York, 1980.

[5] T. Chan, A. Marquina, and P. Mulet. High-order total variation-based image restoration. *SIAM Journal On Scientific Computing*, 22(4):503–516, 2000.

[6] P. Cheeseman, B. Kanefsky, R. Kraft, J. Stutz, and R. Hanson. Super-Resolved Surface Reconstruction From Multiple Images. In *Maximum Entropy and Bayesian Methods: Santa Barbara, California, USA, 1993: Proceedings of the*

Thirteenth International Workshop on Maximum Entropy and Bayesian Methods. Kluwer Academic Pub, Dec. 1994.

[7] M. Elad and A. Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE Transactions on Image Processing*, 6(12):1646–1658, 1997.

[8] Facebook. 10 billion photos, `http://www.facebook.com/note.php?note_id=30695603919`, 2008.

[9] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar. Advances and Challenges in Super-Resolution. *International Journal of Imaging Systems and Technology*, 14(2):47–57, 2004.

[10] W. Feller. *An introduction to probability theory and its applications. Volume I.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Chichester-New York-Brisbane-Toronto-Singapore, 1968.

[11] M. Ghanbari. The cross-search algorithm for motion estimation [image coding]. *IEEE Transactions on Communications*, 38(7):950–953, July 1990.

[12] Google. Get the picture, `http://googleblog.blogspot.com/2005/02/get-picture.html`, 2005.

[13] R. C. Hardie, K. J. Barnard, and E. E. Armstrong. Joint map registration and high-resolution image estimation using a sequence of undersampled images. *IEEE Transactions on Image Processing*, 6(12):1621–1633, Dec. 1997.

[14] Y. He, K.-H. Yap, L. Chen, and L.-P. Chau. A nonlinear least square technique

for simultaneous image registration and super-resolution. *IEEE Transactions on Image Processing*, 16(11):2830–2841, Nov. 2007.

[15] J. Hu and H. Wang. Adaptive total variation based on feature scale. In *ICSP 2004: International Conference on Signal Processing*, 2004.

[16] M. Irani and S. Peleg. Improving resolution by image registration. *CVGIP: Graphical Models and Image Processing*, 53(3):231–239, 1991.

[17] M. Irani and S. Peleg. Motion Analysis for Image Enhancement: Resolution, Occlusion, and Transparency. *Journal of Visual Communication and Image Representation*, 4(4):324–335, 1993.

[18] R. Keys. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(6):1153–1160, Dec 1981.

[19] H. C. Kim. *High resolution image reconstruction from undersampled multiframes.* PhD thesis, Pennsylvania State Univ., Univ. Park, PA, 1994.

[20] S. Kim, N. Bose, and H. Valenzuela. Recursive reconstruction of high resolution image from noisyundersampled multiframes. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(6):1013–1027, 1990.

[21] S. P. Kim and W. Y. Su. Recursive high-resolution reconstruction of blurred multiframe images. *IEEE Transactions on Image Processing*, 2(4):534–539, Oct. 1993.

[22] T. Koga, K. Iinuma, A. Hirano, Y. Iijima, and T. Ishiguro. Motion compensated

interframe coding for video conferencing. In *Proc. Nat. Telecommun. Conf*, volume 5, pages 1–5, 1981.

[23] R. Li, B. Zeng, and M. L. Liou. A new three-step search algorithm for block motion estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 4(4):438–442, Aug. 1994.

[24] L. Liu, E. Feig, I. Center, and Y. Heights. A block-based gradient descent search algorithm for block motionestimation in video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 6(4):419–422, 1996.

[25] S. Mann and R. W. Picard. Virtual bellows: constructing high quality stills from video. In *Proc. ICIP-94. IEEE International Conference Image Processing*, volume 1, pages 363–367, 13–16 Nov. 1994.

[26] H. G. Musmann, P. Pirsch, and H. J. Grallert. Advances in picture coding. *Proceedings of the IEEE*, 73(4):523–548, April 1985.

[27] A. J. Patti, M. Ibrahim Sezan, and A. Murat Tekalp. High-resolution image reconstruction from a low-resolution image sequence in the presence of time-varying motion blur. In *Proc. ICIP-94. IEEE International Conference Image Processing*, volume 1, pages 343–347, 13–16 Nov. 1994.

[28] S. Peleg, D. Keren, and L. Schweitzer. Improving image resolution using subpixel motion. *Pattern Recognition Letters*, 5(3):223–226, 1987.

[29] L. Po and W. Ma. A novel four-step search algorithm for fast block motion estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 6(3):313–317, 1996.

[30] M. Protter, M. Elad, H. Takeda, and P. Milanfar. Generalizing the nonlocal-means to super-resolution reconstruction. *IEEE Transactions on Image Processing*, 18(1):36–51, Jan. 2009.

[31] S. Rhee and M. Kang. Discrete cosine transform based regularized high-resolution image reconstruction algorithm. *Optical Engineering*, 38:1348–1356, Aug. 1999.

[32] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259–268, Nov. 1992.

[33] R. Schultz, L. Meng, and R. Stevenson. Subpixel Motion Estimation for Super-Resolution Image Sequence Enhancement. *Journal of Visual Communication and Image Representation*, 9(1):38–50, 1998.

[34] R. Schultz and R. Stevenson. Improved definition video frame enhancement. In *International Conferenceon Acoustics, Speech, and Signal Processing*, volume 4, pages 2169–2172, 1995.

[35] R. Schultz and R. Stevenson. Extraction of high-resolution frames from video sequences. *IEEE Transactions on Image Processing*, 5(6):996–1011, 1996.

[36] H. Shen, L. Zhang, B. Huang, and P. Li. A map approach for joint motion estimation, segmentation, and super resolution. *IEEE Transactions on Image Processing*, 16(2):479–490, Feb. 2007.

[37] A. Sinha and X. Wu. Fast generalized motion estimation and superresolution. In *Proc. IEEE International Conference on Image Processing ICIP 2007*, volume 5, pages 413–416, Sept. 16 2007–Oct. 19 2007.

[38] C. Srinivas and M. Srinath. Stochastic model-based approach for simultaneous restoration of multiple misregistered images. In *Proceedings of SPIE*, volume 1360, page 1416. SPIE, 1990.

[39] A. Tekalp, M. Ozkan, and M. Sezan. High-resolution image reconstruction from lower-resolution image sequences and space-varying image restoration. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 553–557, 1992.

[40] R. Tsai and T. Huang. Multiframe image restoration and registration. *Advances in Computer Vision and Image Processing*, 1(1):317–339, 1984.

[41] X. Wu, E. U. Barthel, and W. Zhang. Piecewise 2d autoregression for predictive image coding. In *Proc. International Conference on Image Processing ICIP 98*, pages 901–904, 4–7 Oct. 1998.

[42] X. Wu and N. Memon. Context-based, adaptive, lossless image coding. *IEEE Transactions on Communications*, 45(4):437–444, April 1997.

[43] L. Zhang and X. Wu. An edge-guided image interpolation algorithm via directional filtering and data fusion. *IEEE Transactions on Image Processing*, 15(8):2226, 2006.

[44] X. Zhang and X. Wu. Image interpolation by adaptive 2-d autoregressive modeling and soft-decision estimation. *IEEE Transactions on Image Processing*, 17(6):887–896, June 2008.

[45] X. Zhang and X. Wu. Standard-compliant multiple description image coding

by spatial multiplexing and constrained least-squares restoration. In *Proc. IEEE 10th Workshop on Multimedia Signal Processing*, pages 349–354, 8–10 Oct. 2008.

[46] S. Zhu and K.-K. Ma. A new diamond search algorithm for fast block-matching motion estimation. *IEEE Transactions on Image Processing*, 9(2):287–290, Feb. 2000.