Depth-Guided Super-Resolution of 3D Video

by

Özgün Genç

A Thesis Submitted to the

Graduate School of Engineering

in Partial Fulfillment of the Requirements for

the Degree of

Master of Science

in

Electrical-Computer Engineering

Koc University

August 2010

Koc University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Özgün Genç

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

A.Murat Tekalp, Prof. (Advisor)

_____

Serdar Kozat, Asst.Prof.

_____

Mehmet Sayar, Asst.Prof.

Date: _____

ABSTRACT

This thesis investigates the problem of super-resolution (resolution enhancement) in multi-view 3D video.

Although the super-resolution problem has been widely studied for monocular video, work on 3D super-resolution of stereo/multi-view video is rather new. Given recent advances in depth-based stereo/multi-view video representations and compression formats, this thesis proposes a depth-guided super-resolution method for multi-view 3D videos.

Since sub-pixel registration of multiple views is crucial in achieving super-resolution, we employ a parametric registration approach based on affine motion of approximately planar image patches rather than traditional pixel-based stereo matching. The segmentation of planar image patches and their affine registration is guided by depth layer maps.

We also conduct subjective visual evaluation tests of super-resolution in 3D on different stereoscopic displays.

ÖZET

Bu tez 3 Boyutlu (3B) videolar üzerinde çözünürlük iyileştirme problemini inceler.

Çözünürlük iyileştirme problemi monoküler videolar üzerinde daha önce genişçe incelenmiş olsa da stereo/çoklu-görüş videolar üzerinde 3B çözünürlük artırımı ile ilgili çalışmalar oldukça yenidir. Bu tez, derinlik tabanlı stereo/çoklu-görüş video temsil ve sıkıştırma formatlarındaki son gelişmeleri dikkate alarak çoklu-görüş 3B videolar için derinlik yardımlı bir çözünürlük artırma yöntemi önermektedir.

Çözünürlük iyileştirmenin gerçekleştirilebilmesinde çoklu görüşlerin piksel-altı hassasiyette hizalanmasının kritik olması nedeniyle geleneksel piksel tabanlı stereo eşleştirme yerine kabaca düzlemsel görüntü parçalarının ilgin hareketine dayalı bir parametrik hizalama yaklaşımı kullanıyoruz. Düzlemsel görüntü parçalarının bölütlenmesi ve ilgin hizalanmaları derinlik katmanı haritalarıyla yönlendiriliyor.

Bunun yanı sıra, 3 Boyutta çözünürlük iyileştirmenin değişik stereoskopik ekranlar üzerinde öznel değerlendirme testlerini de yürüttük.

iv

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

2D	Two Dimensional
3D	Three Dimensional
3DTV	Three Dimensional Television
SR	Super-resolution
LR	Low Resolution
HR	High Resolution
PSF	Point Spread Function
POCS	Projections onto Convex Sets
LSV	Linear Shift Variant
K	magnification factor
N	number of LR images
n	LR image index
f	desered SR image
g_n	n'th LR observed image
L	number of layers
l	layer index
R	number of sub-regions
r	sub-region index
$I^{n,l,r}$	LR image patch at image n, layer l, sub-region r
I_D^n	depth image of n'th LR image
h_{psf}	point spread function kernel
σ_{psf}	standard-deviation of the point spread function
K_{psf}	point spread function kernel size

Chapter 1

INTRODUCTION

1.1 Motivation

There is growing interest in 3D video formats and services, including 3DTV and 3D mobile video. 3D video formats can be classified as frame-based, including side-by-side and sequential formats, and depth-based including video-plus-depth and multi-view-plus-multi-depth formats. There are several applications where super-resolution (SR) of stereo and multi-view videos is of interest. For example, in the side-by-side transmission format, right and left views are subsampled horizontally by a factor of 2 and composited together in order to utilize the present infrastructure. The composite frames are then split into right and left views which are horizontally upsampled at the receiver. SR methods can be employed during upsampling to prevent loss of resolution. Another application is the display of standard definition multi-view-plus-multi-depth videos on high-definition 3D displays.

SR has been widely studied to obtain high-resolution (HR) frames from a set of low-resolution (LR) images or monocular video. A typical SR method involves two essential

steps: i) sub-pixel accurate registration of low-resolution images/frames on a common HR grid and ii) reconstruction of the HR image using an estimate of the point spread function (PSF) of the LR imaging model. The quality of the resulting HR images strongly depends on the accuracy of the sub-pixel registration and proper estimation of the PSF which relates LR observed pixels to the unknown HR frame. Block-based search or optical-flow based motion estimation are not preferred in SR methods since they are not accurate and stable enough. Parametric motion models, such as affine and perspective models, can achieve the desired sub-pixel accuracy. However, in sequentially recorded images or frames of a monocular video, accurate sub-pixel registration is exceedingly difficult to achieve since local motion of objects in the scene can be too complex to model by a small number of parameters. Combined motion estimation and segmentation methods using multiple parametric motions have been proposed [1; 2].

In a multi-camera setup, assuming the cameras are well synchronized, the scene is virtually static and disparity among different views is due to different camera positions. Parametric models are well suited to model disparity in case of planar scenes, but they fail to model the parallax due to depth variations, if applied to the whole frame. It is possible to represent the scene as a collection of approximately planar layers, where each layer can be registered by a different parametric model. Depth image segmentation can be used to aid layering. Layered disparity registration approaches have been previously proposed for image based rendering, stereo reconstruction and view interpolation [3; 4; 5]. Image patch

similarity across views is exploited in the context of image denoising [6]. We note that depth images can be recorded by time-of-flight cameras or estimated by stereo matching methods.

Although it is desirable to have a shallow focus (small depth of field blurring the background and focusing on the subject) is mostly preferred in conventional 2D cinematography, it is avoided in 3D stereo filmmaking since it gives an effect of individually floating foreground object on a blurry background. Such an effect is generally found annoying by the audience. By using separate PSF for different depth layers, it is possible to widen the depth of field and the SR reconstruction provides a desirable effect of sharp 3D images.

1.2 Problem Formulation

In this work, POCS based SR is applied to the multi-view 3D video sequences having depth-maps to produce a pair of stereo SR images in a higher resolution than the input images. Figure 1 shows a general overview of the whole process.
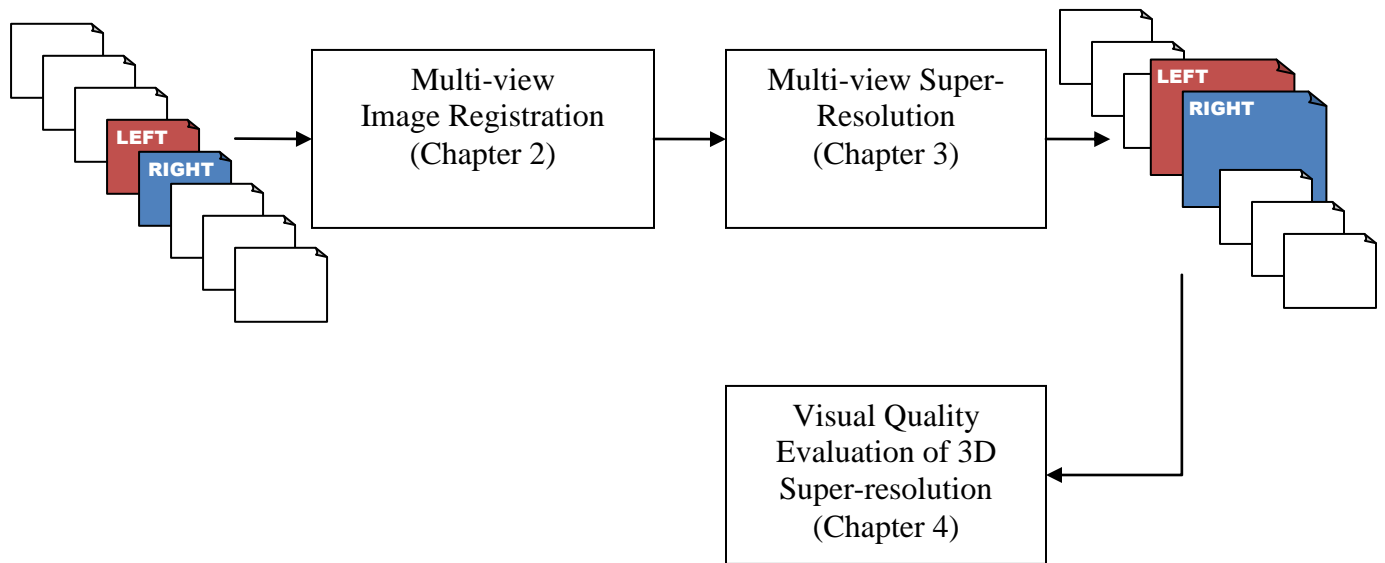
Figure 1: Overview of the System

The aim of the Multi-view Image Registration step is to map the LR pixels from different views at the same time instance onto reference Left and Right views. Sub-pixel accuracy of the registration is a requirement for this step.

In the Multi-view Super-Resolution step we use POCS method to enhance the resolutions of the stereo pair.

A realistic evaluation of the 3D super-resolution requires human inspection since the effect of enhancement should be considered in 3D perception. We conduct subjective tests to evaluate the performance of the 3D super-resolution.

1.3 Contribution

In this thesis, we propose a depth-image-guided, layered multi-view registration approach together with the Projection onto Convex Sets (POCS) framework in order to achieve multi-view SR. The output can be monocular, stereo or multi-view HR images or video. We also performed perceptual quality evaluation tests to compare SR stereo images with linearly interpolated stereo images. The main contributions of this paper are: i) a depth-guided multi-view image registration method, ii) evaluation of 3D SR results in two different 3D displays.

The thesis is organized as follows: The method for depth-guided registration of image patches from multiple views is described in Section 2. Section 3 describes a POCS based SR reconstruction method. Section 4 discusses perceptual evaluation of SR in 3D. Experimental results are presented in Section 5. Finally, Section 6 draws some conclusions.

Chapter 2

MULTI-VIEW IMAGE REGISTRATION

2.1 Introduction

Image registration is the process of overlaying two or more images (image regions) of the same scene on a common coordinate system. It is an essential step in multi-frame image restoration where information from multiple frames is combined. It is also a very important step in the solution of the SR problem.

Although there are some SR techniques proposed where the registration is done simultaneously with the SR enhancement [7; 8], image registration is mostly a separate initial step in the whole process. In these simultaneous methods the registration parameters (mostly shifts and rotation) are incorporated in the optimization process. The results of simultaneous registration and enhancement can be promising since the registration is done considering the enhancing resolutions and the two processes are inter-dependant. Such a perspective, however, is more suitable for stochastic SR methods and on images with single global motion. A simultaneous registration is inefficient for pixel-based POCS framework.

Image registration is the separate first step in our super-resolution method as well (Figure 1). The overall aim of this step is to register all the LR pixels onto a common grid in HR. Since the destination image is not available in HR at the registration step, the motion estimation is done at the LR level and the motion is expanded by the magnification factor K.

Registration is the most critical step in multi-frame super-resolution. This is because the extra information needed to reconstruct high frequencies is contained in the sub-pixel displacements between the images. Such a sub-pixel displacement corresponds to a displacement in the sampling lattice; hence the unknown original signal is sampled and observed in a new way. Displacements with full pixel lengths on the other hand, provide no new information and have no use for super-resolution purposes.

This essential role of sub-pixel displacements imposes a limitation to the registration method to be used in super-resolution: Registration must be sub-pixel accurate.

The second requirement is on the robustness of the registration algorithm. Misregistered pixels introduce unacceptable artifacts to the resulting image and make the super-resolution infeasible.

Throughout this thesis the term motion estimation is used to refer to the same process of estimating the correspondence between the pixels on the source and the destination images. Since the images to be registered are the different camera views taken at the same time

instance, the registration problem can also be stated as the stereo-matching problem as commonly used in computer vision.

Yet, in literature motion estimation usually refers to finding the motion of individual pixels in consecutive video frames, whereas the common stereo-matching algorithms deal with establishing one-to-one pixel relationships between the pixels from the left and right views.

Stereo vision is commonly used in computer vision and multi-view scene reconstruction context. Capel et al. [9] applies the computer vision techniques to the SR problem. As a general framework, some easily detectable feature points are matched for the left and right views, and these correspondences are fed to a model fitting algorithm such as RANSAC to estimate a Homography Matrix that is in accordance with the most of the feature correspondences by omitting outliers. Such a Homography Matrix defines a projective transformation between the left and right views. This transformation can then be refined using a search-based dense matching algorithm to find the actual pixel-by-pixel mappings [10]. With the additional knowledge of the intrinsic camera calibration parameters such as the focal length and lens distortions, it is also possible to map the 2D pixels to 3D scene locations.

Although such a 3D-vision framework seems to solve the multi-view registration problem of super-resolution, it is unnecessarily complex and still suboptimal in satisfying our sub-pixel registration requirements.

We instead propose a video processing based solution without dealing with the 3D scene geometry or epipolar geometry. We instead employ common and more standardized motion estimation and video processing algorithms and we consider the problem as a 2D warping and 2D reconstruction process. Our registration method requires no calibration at all, and is in accordance with the multi-view-plus-multi-depth format.

Another common way of motion estimation in video sequences is using block-matching methods, where the pixel blocks are searched and registered separately on the target image. However, the aperture problem is apparent here and the matches can be ambiguous (i.e. many similar blocks are possible on the target image). Such an ambiguity may not be a problem when the motion estimation is used for coding purposes; however, it is definitely unacceptable in SR reconstruction. Besides its sub-pixel accuracy is limited and is dependent on an interpolation operation.

Although Baker and Kanade [11] have combined the optical-flow algorithm simultaneously into SR reconstruction, using optical-flow in SR is found infeasible by other authors [12].

These block or pixel based methods are more suitable when there are non-rigid motions (e.g. changing face expressions). In our multi-view setup, all the images are taken at the same time instance and all the motions to be estimated are rigid even if there are non-rigid motions in the time axis.

2.2 Patch-based Motion Estimation Using Depth Segmentation

This section addresses mapping individual LR pixels onto a reference HR grid. This reference HR grid is the upsampled version of the LR view to be super-resolved. All the LR images are registered to the reference LR view. The patch-based registration parameters are used to synthesize pixel mapping from each LR pixel to the reference image, and these mappings are expanded by the magnification factor K to map onto the reference HR grid. Each LR pixel is associated with an indicator in the verification map, indicating the reliability of its mapping.

We propose to employ the depth image as a constraint to register similar patches in multi-view images with sub-pixel accuracy. We assume that the depth images are available [13]. Actually, the depth information is considered as a part of the video format.

Depth-map generation is considered as an error prone task. The depth values can be inaccurate and noisy. In our method, however, the precise accuracy of the depth images is not essential since we use depth images only to guide layering. The depth-maps are used to segment images into image patches having approximately planar motion under stereo camera transformations.

Our approach for registration is to model each image frame (viewpoint) as a collection of L layers each composed of multiple patches, and to assign a different parametric motion model to each image patch. A typical multi-camera setup has a linear or a curved arrangement, and our approach can provide an accurate and reliable motion model for most

image regions. Since the POCS framework, used in SR reconstruction, allows us to project onto only LR pixels that are accurately registered, the registration method does not need to perfectly register all pixels.

Image layers, that are approximately planar, are extracted by a segmentation of the depth images. The binary mask $I_M^{n,l}$ indicating layer l of frame n is defined as:

$$I_M^{n,l}(i,j) = \begin{cases} 1, & T^{n,l} \leq I_D^n(i,j) \leq T^{n,l+1} \\ 0, & otherwise, \end{cases} \tag{1}$$

where I_D^n is the depth image and $T^{n,l}$ and $T^{n,l+1}$ are upper and lower depth threshold values for layer l of image n. These threshold values are set considering abrupt depth changes at the object boundaries and the number of required pixels for stable registration.

Since a single affine model may not provide a perfect fit to a whole image layer due to small depth variations, we propose using a piecewise affine model within each layer, where we further divide layers into R sub-regions, called patches, and fit different affine models to each of these patches. Figure 2 shows image patches generated by using L=11 depth levels and R=3 horizontal sub-regions. Here, only simple horizontal sub-divisions are found sufficient since dominant camera motion for this dataset is horizontal.
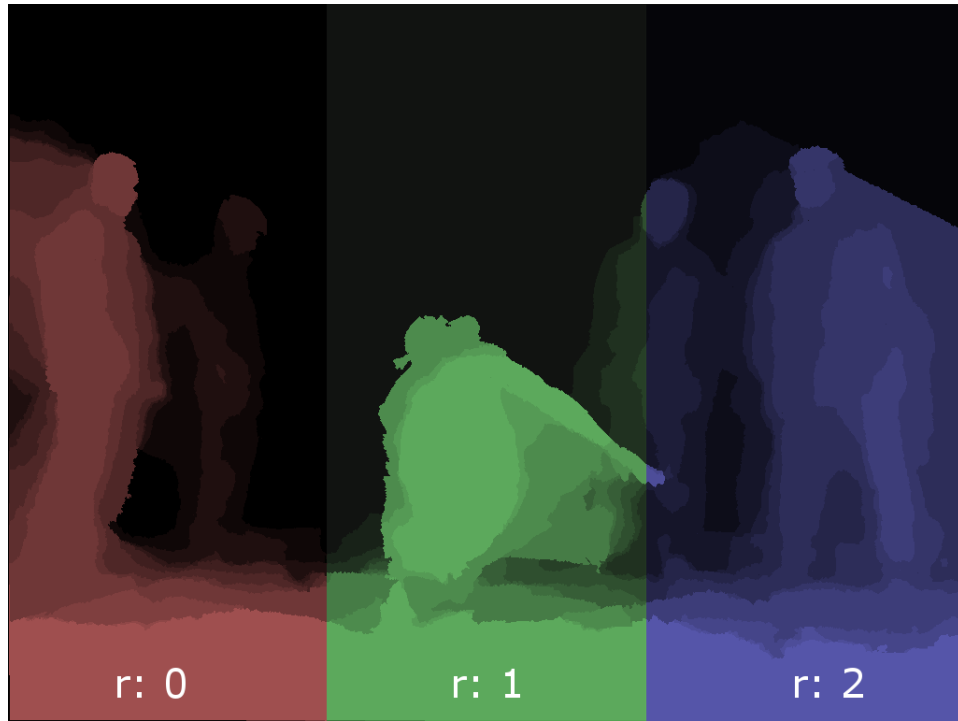
Figure 2 - Image patches: Each image patch segment is indicated with a different color/intensity

Affine registration parameters from each one of these image patches to the reference frame are estimated by providing patch masks $I_M^{n,l,r}$ ($r \in R$) as input image supports to the Motion2D registration algorithm [14]. Motion2D is a M-estimator based parametric registration library that can operate on hierarchical manner.

The parametric registration algorithm iteratively fits a parametric motion model to provide motion vectors per pixel in the most general form of:

$$\begin{bmatrix} u(x,y) \\ v(x,y) \end{bmatrix} = \begin{pmatrix} c1 \\ c2 \end{pmatrix} + \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} q_1 & q_2 & q_3 \\ q_4 & q_5 & q_6 \end{pmatrix} \cdot \begin{pmatrix} x^2 \\ xy \\ y^2 \end{pmatrix}, \qquad (2)$$

where x and y are the pixel coordinates on the source LR image; $u(x,y)$ and $y(x,y)$ are the horizontal and vertical components of the motion vector at the point (x,y); c_1 and c_2 are the horizontal and vertical translations, a_i are linear transformation parameters correspond to rotation and shear motions; and q_i are the quadratic parameters representing the linearization of the perspective motions.

This 12-parameter motion model represents a wide range of transformations. However the quadratic q parameters are found to be zero in our practical experiments and we found it sufficient to concentrate on the 6 parameter affine model.

Mappings from each LR pixel onto the HR grid are synthesized from these parameters.

The difference between the LR pixel value on the source image and the value of the pixel at its mapped coordinate provides a confidence weight about the correctness of that pixel registration:

$$I_W^n(x,y) = I^{ref}(x + u(x,y), y + v(x,y)) - I^n(x,y) + \zeta_n \tag{3}$$

where ζ_n is the global intensity change between the I^{ref} (reference LR image) and I^n (n^{th} LR image) as computed by the Motion2D algorithm.

We construct a verification map V^n for n^{th} LR image using these registration weights along with depth and reconstruction cues such that:

$$V^n(x,y) = \begin{cases} 0, & I_W^n(x,y) < T_w \\ 0, & |I_D^{ref}(x + u(x,y), y + v(x,y)) - I_D^n(x,y)| < T_D \\ 1, & otherwise \end{cases} \tag{4}$$

In these equations T_w is the weights threshold and T_D is the depth threshold. $V^n(x, y)$ is also set to zero during the POCS iterations when the residue at LR pixel (x, y) is larger than a predefined reconstruction threshold T_R.

Registration of an image patch is illustrated in Figure 3.

Figure 3 - Registration of a single image patch onto the reference image: (a) unregistered (b) registered with the motion parameters of the global layer estimation (i.e. R=3) (c) refined registration with motion estimation at the image patch (R=3)

Note that restricting the target image support to the pixels only with similar depth values as the input patch is also possible (Figure 4). In this way it is possible to further guide the registration of the image patch using depth-maps.

Several parameter initialization strategies can be adopted to ease and to guide the patch registration. These strategies make sense since the chosen parametric registration method is designed to be operated for small motions. The multi-view geometry, on the contrary, causes to large displacements in nature. Although the registration algorithm tries to overcome this problem by operating in hierarchical (i.e. image pyramids based, multi-resolution) manner, we may improve the registration performance by firstly estimating a global motion for the layer, and then committing refining estimations at all its R sub-regions (i.e. image patches at that depth layer). We can also compare the registration weights of these layer global and local image-patch registrations at pixel level to determine the best motion parameters per LR pixel.

If the camera calibration parameters are available, they can also be used as initial guesses for the registration. To do this, we compute the affine parameters for each of the image patches using the calibration matrices and feed these affine parameters to the registration algorithm as initial estimates. This approach gives better results than registering pixels using only the calibration matrices. This better performance is expected considering that the depth images are noisy in general.

Figure 4 – Registration with destination support: (a) unregistered patch, (b) destination support, (c) registered

To compute the affine parameters from the calibration matrices we assumed three points representing the whole patch, mapped them onto the target frame using the known calibration matrices and found the affine transformation in between. The x and y coordinates of these three points are around the center of mass of the depth image patch and their depth is the average depth value of the image patch (Figure 5).
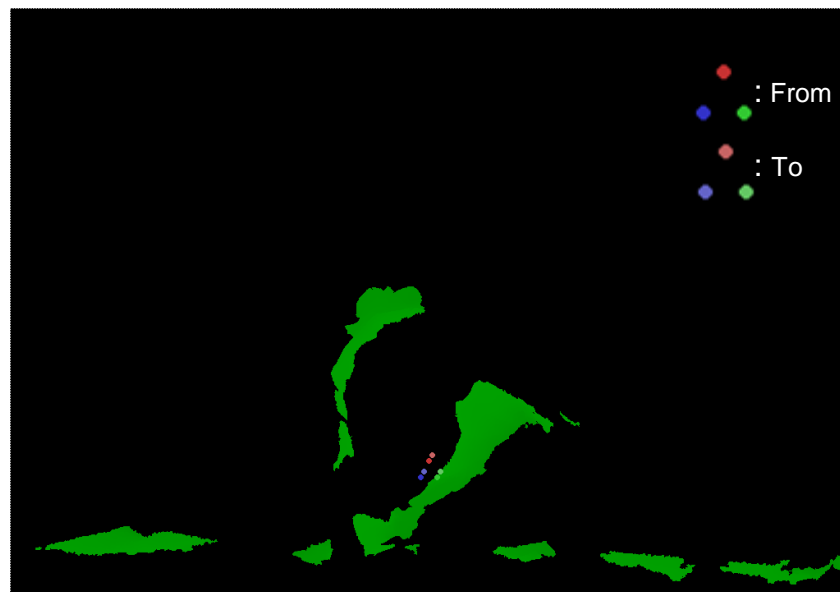
Figure 5: Affine motion of a layer patch derived from calibration parameters

Chapter 3

MULTI-VIEW SUPER-RESOLUTION

3.1 Introduction

Super-resolution is the process of increasing the spatial resolution of a given image. It differs from ordinary image zooming or interpolation techniques in the way that it recovers the high frequency components existing in the original scene but disappeared during the imaging process.

When it is put in this way it is evident that the super-resolution is an ill-posed problem. A low resolution image does not provide the necessary information for high frequency recovery on its own. SR algorithms are broadly divided into two main groups according to their solution to this problem: (i) Learning based SR algorithms and (ii) Reconstruction based SR algorithms.

Learning-based methods [15] attack the problem with machine learning solutions. In a typical application of this method, a training dataset of high-resolution image patches and their low-resolution counterparts is used. Given a low-resolution input image, corresponding high-resolution image patches are estimated for each LR pixel and are

combined together to obtain a super-resolved output image in HR. Such an approach is feasible only when the content of the images are either known (known scene) or are limited with a certain class of input images (e.g. face images, text etc.)

Reconstruction-based methods, on the other hand, can be applied for a general set of input images and the SR algorithm used in this thesis is of this kind. Reconstruction-based super-resolution techniques have their roots in Sampling Theory and in the well studied field of Image Restoration.

In a general image restoration problem, the input image to be restored is treated as the output of an image formation/degradation system, where the input of the system is the real scene or a hypothetical unknown HR image. The problem, then, reduces to modeling the degrading image formation process. We call this model as the Observation Model (see Section 3.2 for the discussion of the Observation Model used in our work). The restoration is achieved by applying the inverted degrading system (W^+) to the observed image. This kind of procedures are used extensively for many applications including blur and noise removal.

For the super-resolution problem, however, the problem is ill-posed by nature. An LR image can be obtained by sub-sampling infinitely many different HR images and W is highly underdetermined. This property of the process makes single image super-resolution an infeasible option for reconstruction-based approach.

Instead of using a single LR image, a series of LR images of the same scene are used in almost all reconstruction-based super-resolution methods. By this way, it is possible to make use of the information exposed by the sub-pixel displaced multiple observations of the same LR pixel. This observation is the key idea behind the reconstruction-based super-resolution.

An early super-resolution technique [16] illustrates this concept and summarizes the steps of the SR reconstruction process as:

(i)   Image Registration (covered in Chapter 2)

(ii)   Interpolation

(iii)   Restoration

These steps can be separately or simultaneously handled.

Elaborating on this framework many super-resolution algorithms are proposed over the last two decades. The basic differences among these algorithms are mainly in their way to regularize and solve the system of linear equations:

$$\underline{y} = W\underline{x} \tag{5}$$

Probabilistic methods such as ML (Maximum Likelihood), MAP (Maximum a Posteriori) [17] and MRF (Markov Random Fields) concentrate on the regularization and iteratively solution of this system of linear equations.

Set theoretical methods including POCS (Projections onto Convex Sets), IBP (Iterative Back-Projections) and Landweber Iterations on the other hand converge to a solution in the solution set in an iterative manner by pixel operations.

The super-resolution method proposed in this thesis is using the POCS framework as its main super-resolution algorithm. Section 3.3 discusses POCS and the proposed super-resolution approach in more details.

3.1.1 Observation Model

Determining the observation model is an essential part of many image reconstruction applications including super-resolution. This model defines the relationship between the degraded observed signal and the desired one.

In SR purposes we are interested in modeling the degradations such as the optical blur, motion blur, sensor noise sub-sampling and aliasing.

In many SR methods, LR images are considered as the warped, blurred, sub-sampled versions of the HR image, with additional noise.

Figure 6 shows a block diagram that is consistent with our method, with M_i being matrices representing the warping, B is the blur matrix and D is the decimation matrix.

Although slightly different observation models are proposed for different super-resolution algorithms in the literature, most of them can be reduced into the form:

$$\underline{y} = W\underline{x} + n, \qquad\qquad (6)$$

$\underline{x}$ and $\underline{y}$ being the vectors formed by the lexicographically ordered pixel values of f and g, respectively.
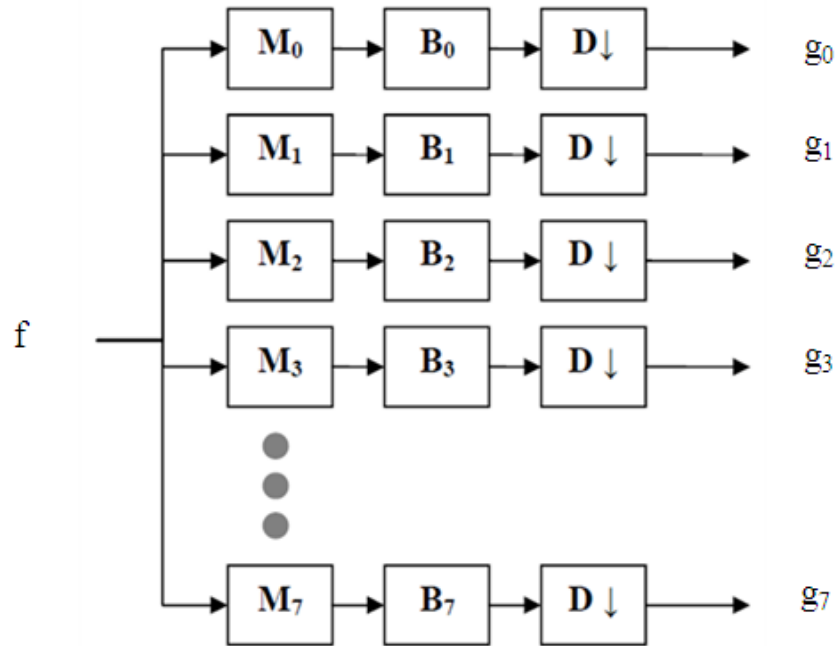
Figure 6 – Block diagram of the Observation Model

This form implies that the observed LR pixels are formed by the weighted sum of some of the HR pixels in the desired HR image. Establishing this relationship between the desired HR pixels and their observations in LR is the most challenging part in the super-resolution process and is inclined to have a high degree of errors since it is itself an estimation process involving blur identification and motion estimation.

Once this relationship is established, the system can be regularized and be solved using traditional numerical methods. In this work, however, we use the POCS approach to let the algorithm iteratively converge to the solution and an explicit construction of W matrix is not needed. Instead, we treat each observed LR pixel separately and find the corresponding HR pixels to be linearly combined. The theory is that, the LR observation should be simulated using these HR pixels, with an error related to the noise process.

Image registration step provides the necessary motion parameters to map individual LR
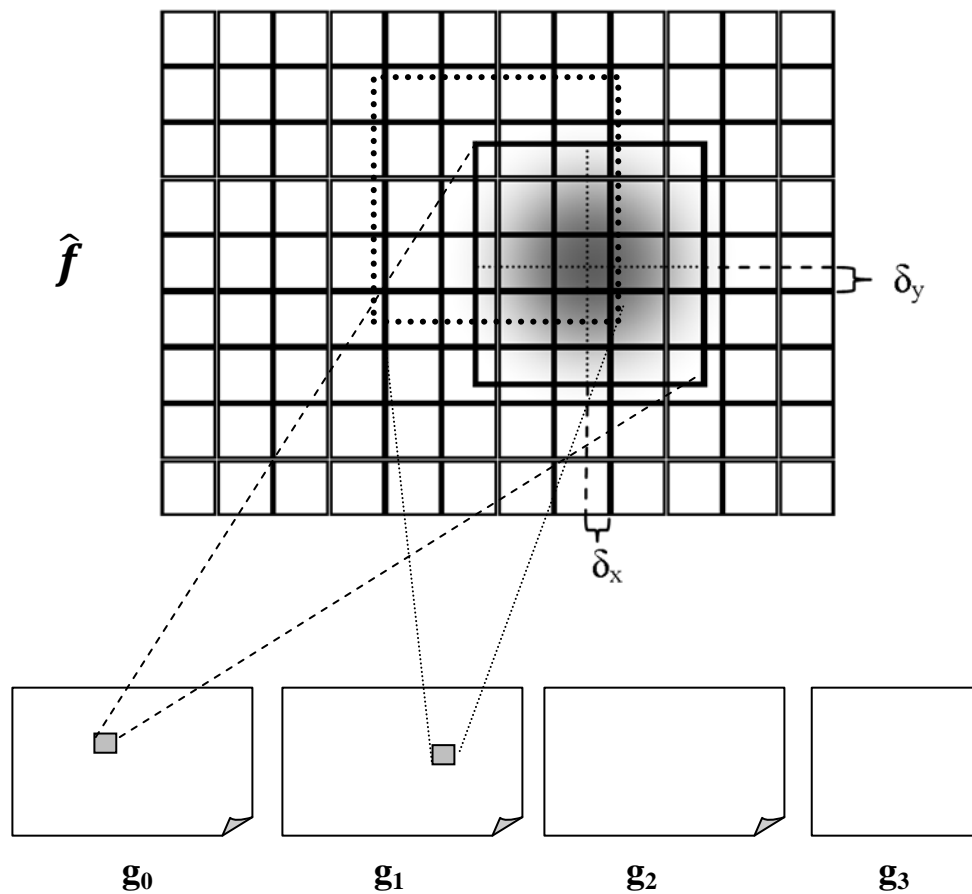
Figure 7 – LR pixel generation from HR grid using PSF

pixels onto the common HR grid on the reference view. Since the estimated motion is a parametric motion, this mapped point has a non-integer coordinate on the reference HR grid. We accept this point as the origin of the PSF and we obtain the 2D discrete PSF coefficients by sampling the continuous PSF generating function at integer locations of the HR grid (HR sampling lattice). (see Figure 7)

In our work, this PSF generating function is the 2D Gaussian function, which is a suitable choice to model the blurring and downsampling effect. Since the function is circularly symmetric translational and rotational camera motions it preserves its shape.

3.1.2 Overview of the POCS Algorithm

POCS (Projections onto Constraint Sets) is a set theoretical optimization method that has various applications in image processing. This method forms the core of the SR reconstruction algorithm.

POCS method is firstly adapted to the solution of super-resolution problem by [18]. Patti et al. extends the method to cover a more complete model of image formation accounting noise and various LSV blurs including the blurring due to non-zero aperture-time [19].

POCS method restricts the solution with convex constraint sets C_i and iteratively converges to a solution in the intersection set $C_s = \bigcap_{\forall i} C_i$.

In POCS SR reconstruction each LR pixel defines a reconstruction constraint set:

$$C_{x,y;n} = \left\{ f(k,l) : \left| r_n^{(f)}(x,y) \right| \leq \delta_0 \right\} \tag{7}$$

where δ_0 is a confidence bound mostly chosen to be related with the noise process. The residue $r_n^{(f)}$ can be defined as:

$$r_n^{(f)}(x,y) = g_i(x,y) - \sum_{k=0}^{N-1}\sum_{l=0}^{N-1} f(k,l) h_{psf}(x,y;k,l). \tag{8}$$

The projection operator P projects the current estimate f to the constraint set so that the estimate becomes consistent with the constraint:

$$P_{x,y;n}[f] = \begin{cases} f(k,l) + \dfrac{r_i^f(x,y) - \delta_0}{\sum_o \sum_p h_n^2(x,y;o,p)} h_{psf}(x,y;k,l) & if\ r_n^{(f)}(x,y) > \delta_0 \\[2mm] f(k,l) & if\ -\delta_0 > r_n^{(f)}(x,y) > \delta_0 \\[2mm] f(k,l) + \dfrac{r_n^f(x,y) + \delta_0}{\sum_o \sum_p h_n^2(x,y;o,p)} h_{psf}(x,y;k,l) & if\ r_n^{(f)}(x,y) < -\delta_0 \end{cases} \tag{9}$$

Since POCS method is slow in convergence we use a simplified version of the projection operator:

$$P_{x,y;n}[f] = \begin{cases} f(k,l) + r_i^f(x,y).h_{psf}(x,y;k,l) & if\ V(x,y) = 1 \\ f(k,l) & if\ V(x,y) = 0 \end{cases} \tag{10}$$

The solution is achieved by successively project onto different constraint sets until convergence is achieved.

The advantage of the POCS method is at its flexible control on the by pixel basis. It allows effective integration of various data such as region of interests, segmentation maps, validity maps or space-varying blur functions and space varying motion segments. The

most important advantage is that its low memory requirement since keeping huge matrices is not needed in contrast of the other methods.

3.2 Super-resolution Image Reconstruction from Multiple Views

This chapter discusses reconstruction of a pair of HR left and right views from N views taken at the same time instance. The number of required views depends on the amount of magnification to be achieved. Typically, more than $N = 2^K$ views are required to achieve KxK resolution increase. We employ the POCS method for SR reconstruction. The main advantage of the POCS method is its flexibility in using independent constraint sets at each pixel; hence, its ease of adaptation to accuracy of patch registration.

We use each LR observation pixel as a constraint as long as we can find a valid motion between this pixel and the target SR frame. Each valid constraint pixel should be able to be simulated from the estimated target SR image. This simulated value is the linear combination of the HR pixels around the location where the LR pixel is mapped on the target SR frame. Coefficients of this linear combination are also known as the PSF (Point Spread Function).

The SR algorithm simultaneously simulates LR pixels and back-projects the residual (difference between the simulated and observed value of the LR pixel value) onto the target frame using the same PSF. LR pixels with simulation errors (residuals) larger than a given threshold value are omitted since they are accepted as unable to be simulated with the

estimated registration values. This check is useful to detect most of the registration errors and to omit pixels belonging to covered regions.

It is also possible to use the depth information in order to further check the validity of this relationship between the LR and HR pixels. Ideally an LR observation pixel should be simulated from HR pixels with similar depth values when it belongs to a non-boundary region of a smooth surface. Although using such a constraint requires exact and reliable depth maps at HR level, we used a simple depth check at the LR level.

PSF Estimation:

The PSF h_{psf} determines the relationship between the HR pixels and the LR pixels. The effect of the sensor geometry, focus blur and relative motion along with the sub-sampling operation can be modeled as a convolution operation with a linear shift variant PSF [19]. Using a parametric function reduces the problem to parameter estimation. Two-dimensional Gaussian is a good choice to model the PSF. We considered only a space-invariant blur caused by the imaging optics. Since all low resolution images are taken at the same time instance, motion blur between the frames is no longer needed to be accounted.

Two-dimensional Gaussian kernel is determined by two parameters: standard deviation σ_{psf} and the kernel support K_{psf}. Among these two parameters σ_{psf} determines the main characteristic of the function and has the primary importance. Once the σ_{psf} parameter is determined K_{psf} can be set to an appropriate value such that the effect of truncation is minimal. Otherwise, truncating the Gaussian function may introduce zero crossings in its

Fourier transform which would result in ringing artifacts in the SR images. Figure 8 shows some PSF and their frequency responses to illustrate this concept.

To find good values of the parameters we generate a series of h_{psf} kernels using σ_{psf} and K_{psf} values varying over a given range. The basic assumption is that the resulting image will be the most plausible when the POCS algorithm is run with the correct parameters.

During the experiments we observed that using a too small σ_{psf} has a negative effect of damaging the edges by visible mis-registered pixels. Using a too large σ_{psf} on the other hand deforms and blurs the image by the exaggerating ringing artifacts. The results were still acceptable over a range of σ_{psf} values which implies that the precise selection of the parameters is not critical. Actual SR results for various PSF parameters are presented in Chapter 5.
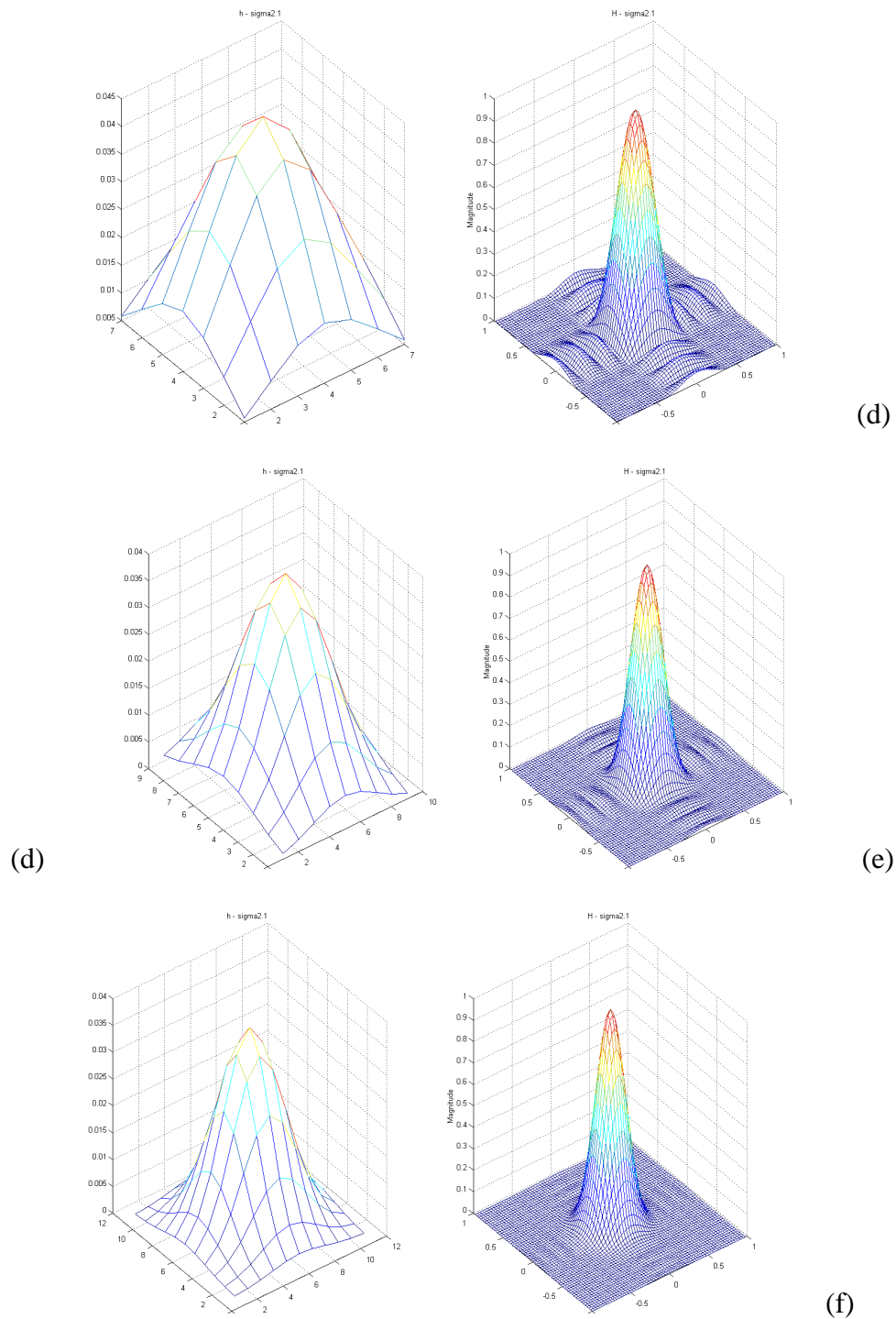
(a)

(b)

(c)
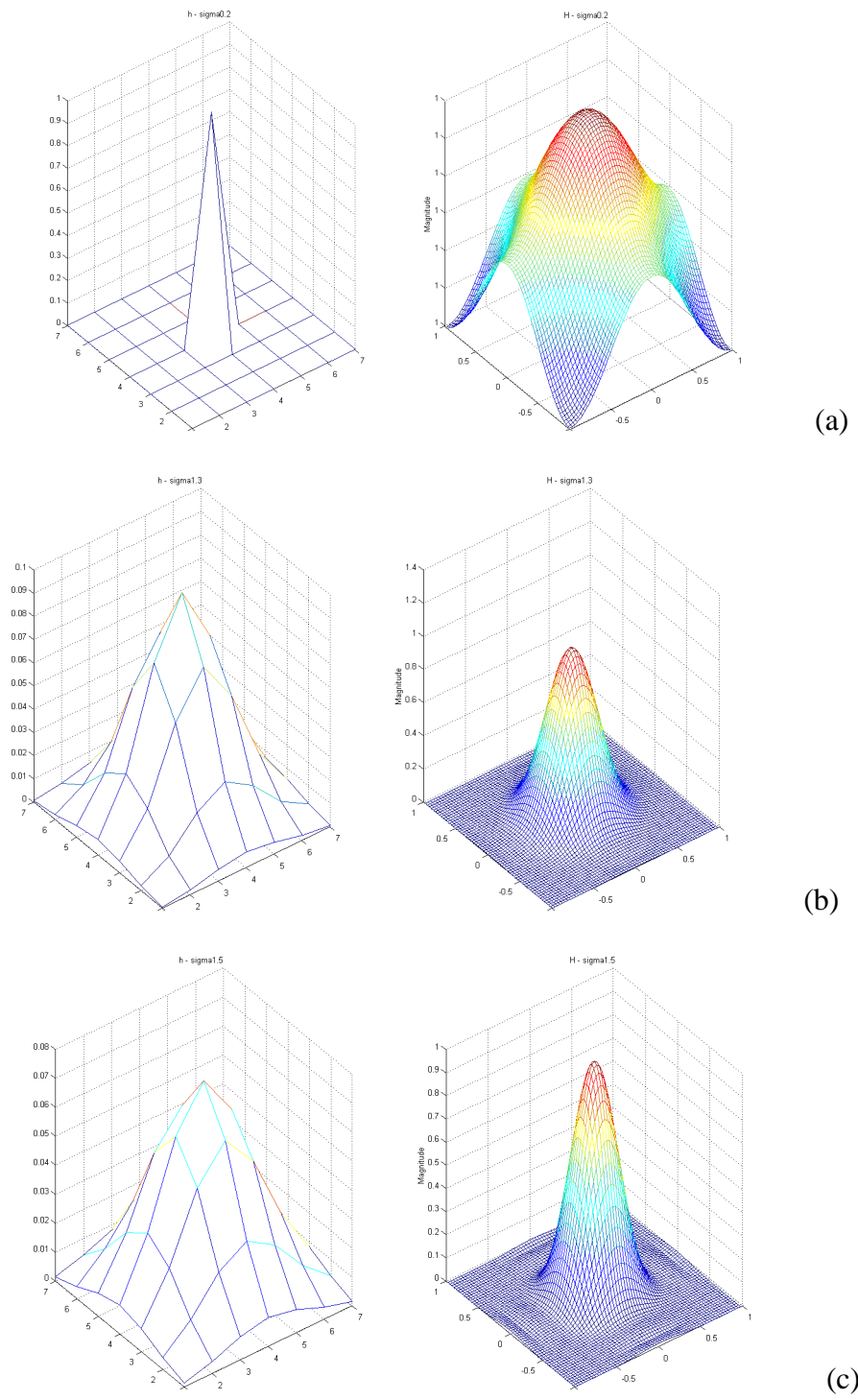
Figure 8 – Various 2D Gaussian PSFs: (a) $K_{psf} = 7, \sigma_{psf} = 0.1, (b)\ K_{psf} = 7, \sigma_{psf} = 1.3,\ (c)K_{psf} = 7, \sigma_{psf} = 1.5, (d)K_{psf} = 7, \sigma_{psf} = 2.1, (e)K_{psf} = 9, \sigma_{psf} = 2.1, (f)K_{psf} = 11, \sigma_{psf} = 2.1.$

Extension to Temporal SR:

In our method the super-resolution is achieved using different camera views at the same time instance, which means that the additional information from consecutive video frames can still be utilized. We can propose that using a reconstruction-based super-resolution algorithm is more suitable for making use of the multi-view information, whereas the temporal axis is more suitable for a learning based super-resolution approach. As for POCS framework, a super-resolved video frame can be used as the initial estimate for the super-resolution of the next frame, after warped using block-based inter-frame motion estimation. Note that such a warping does not require sub-pixel accuracy in contrast to the approach of fusing consecutive video frames. The effect of adding such a variation to the POCS reconstruction will be a decrease in the convergence time.

For a learning based or Bayesian (i.e. MAP or MRF) super-resolution framework, on the other hand, the previous SR frames can be used to determine the scene specific priors.

It is also possible to directly map the LR pixels from the adjacent time instances. For an 8-camera system, 24 frames of the same scene are available in only ±1 shutter-period time range. Such an amount of redundant data is suitable for a 4 by 4 SR magnification.

Chapter 4

VISUAL QUALITY EVALUATION OF 3D SUPER-RESOLUTION

4.1 Introduction

Quality evaluation of 3 dimensionally perceived images are a relatively new task and works on this subject are limited. There is no standardized way of measuring the quality of a 3D image and the relations between the visual quality the quantitative metrics are still obscure. The consequence of this observation is that the true performance of the SR in 3D requires subjective evaluation.

In this chapter of the thesis, we describe our method of subjectively measuring the effect of super-resolution on 3D video.

4.2 Visual Quality Evaluation of 3D Super-Resolution

We conducted two subjective tests for analyzing the performance of the super-resolution on two different 3D displays to investigate the effect of the display type, using Double-Stimulus Continuous Quality-Scale (DSCQS) method [20]. 8 assessors are selected for the tests. The first setup consist of a pair of Sharp MB-70X projectors, a silver dielectric screen, polarized filter glasses and a PC to drive the projectors. The second 3D display is a Sharp AL3DU laptop that uses parallax barrier.

The assessors are expected to grade test images generated with the proposed super-resolution method, in comparison with the interpolated (for Set 1 results) or original (for Set 2 results) pairs. Before grading, each test image and its original (or interpolated) version are displayed twice in a random order.

The grades are then evaluated and converted to a standardized test score so that the different algorithms can be compared.

For further information on the test setup and the evaluation procedure, please refer to [21].

Chapter 5

RESULTS

5.1 Introduction

We conducted our experiments on the Breakdancing sequence [22]. This sequence contains 8 camera views with depth maps generated by a stereo matching algorithm. The cameras are arranged along a slightly curved line (Figure 9) so the inter-camera transformations are mainly translational with a small amount of rotation. Such camera arrangements are common for multi-viewpoint 3D video contents since they imitate the viewpoints of the audience in front of a 3D screen and provide the parallax effect which is essential for 3D perception. All the cameras are static and the camera calibration matrices are also available. We use the 8 views from a single time instance (third frame in the sequence) to reconstruct a pair of super-resolved images. We chose the 4th camera as the left view and the 3rd camera as the right view of the stereo pair.

In Section 5.2 we present some results regarding the performance of the multi-view image registration methods described in Chapter 2. Section 5.3 is on the evaluation of the multi-view super-resolution. Quantitative evaluation results are given in Section 5.3.1 and the subjective results obtained using the test procedure described in Chapter 4 are provided in Section 5.3.2.

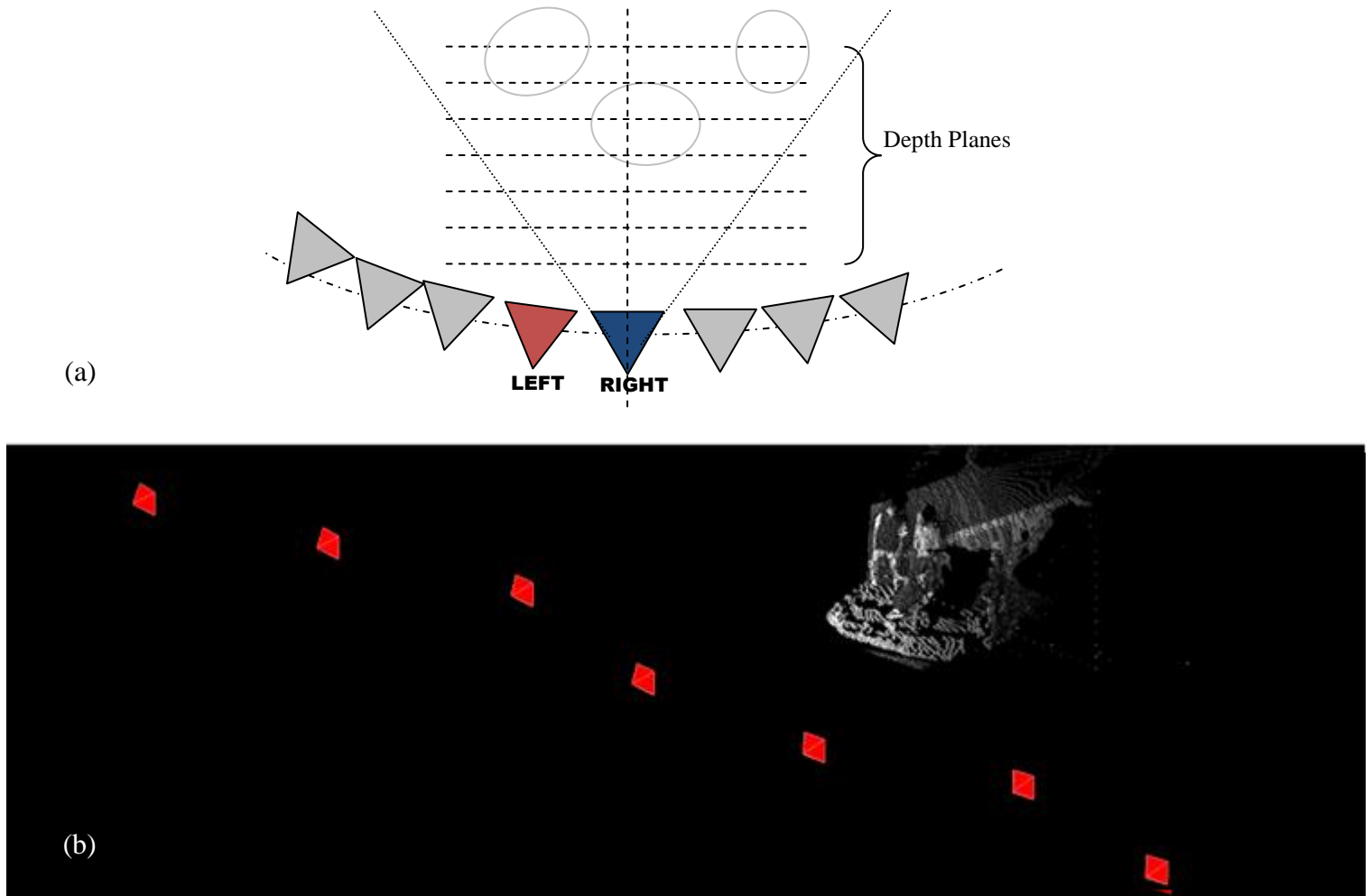Figure 9 – Camera Arrangements: (a) Top-view, (b) Actual camera locations and the
3D point cloud representation of the scene from a perspective view

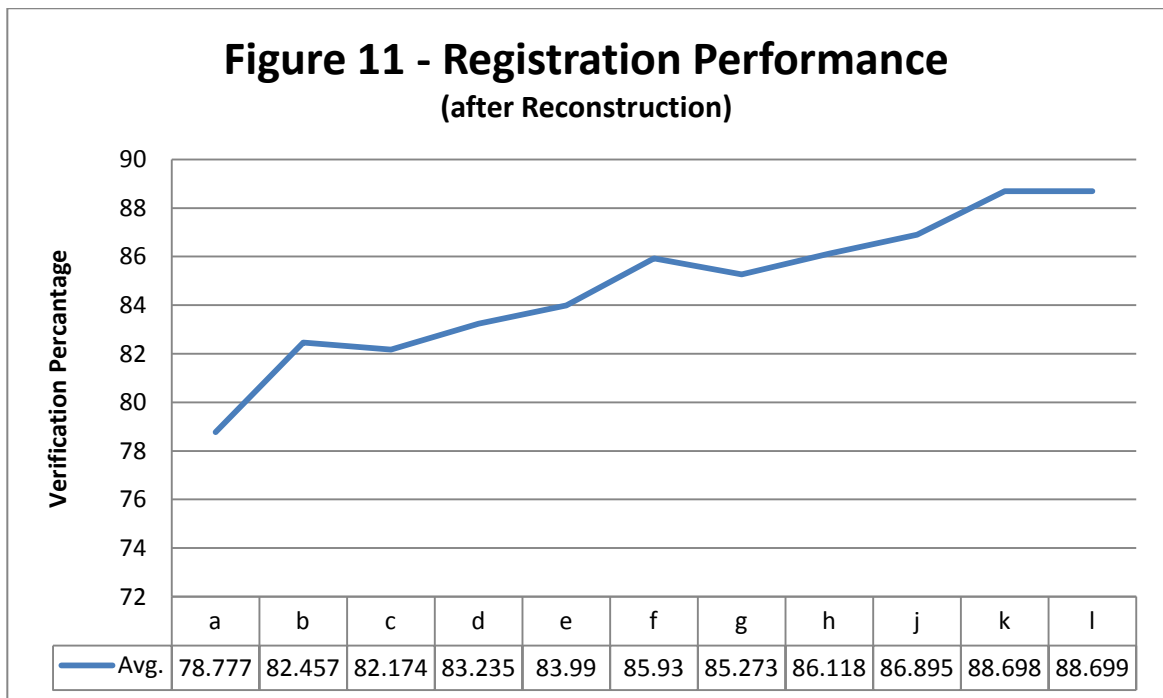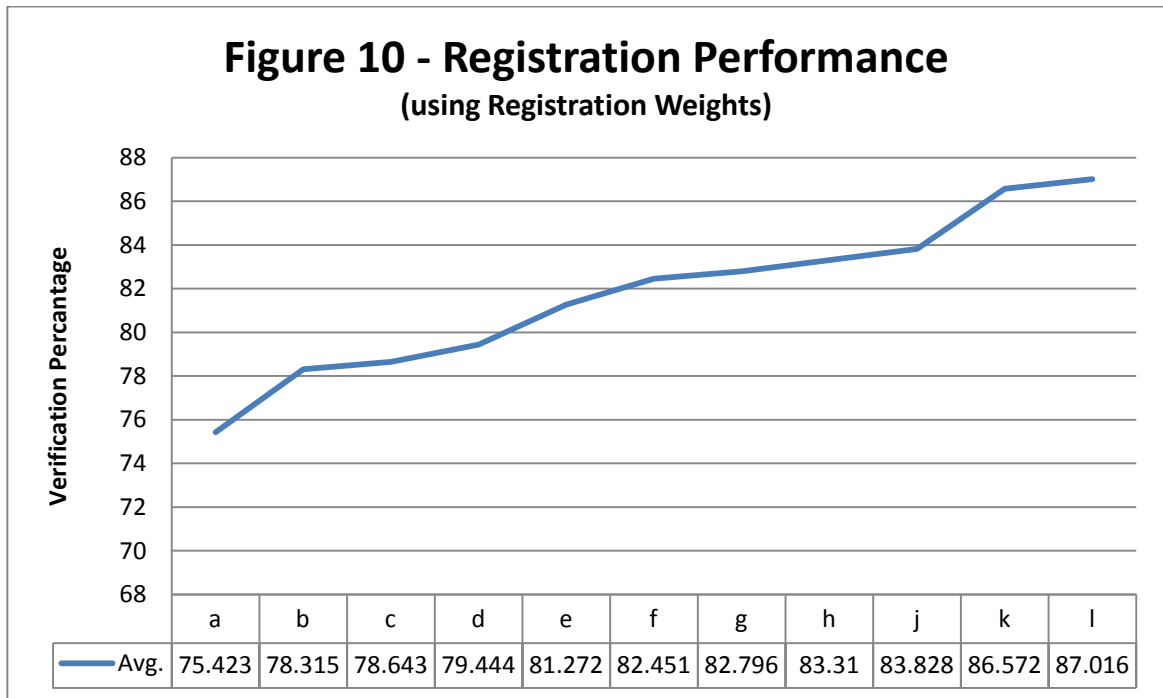## 5.2 Results of Multi-view Image Registration

Although the ground-truth motion is not available, the performance of the proposed

image registration method can be evaluated with the help of the registration weights I_W^n and
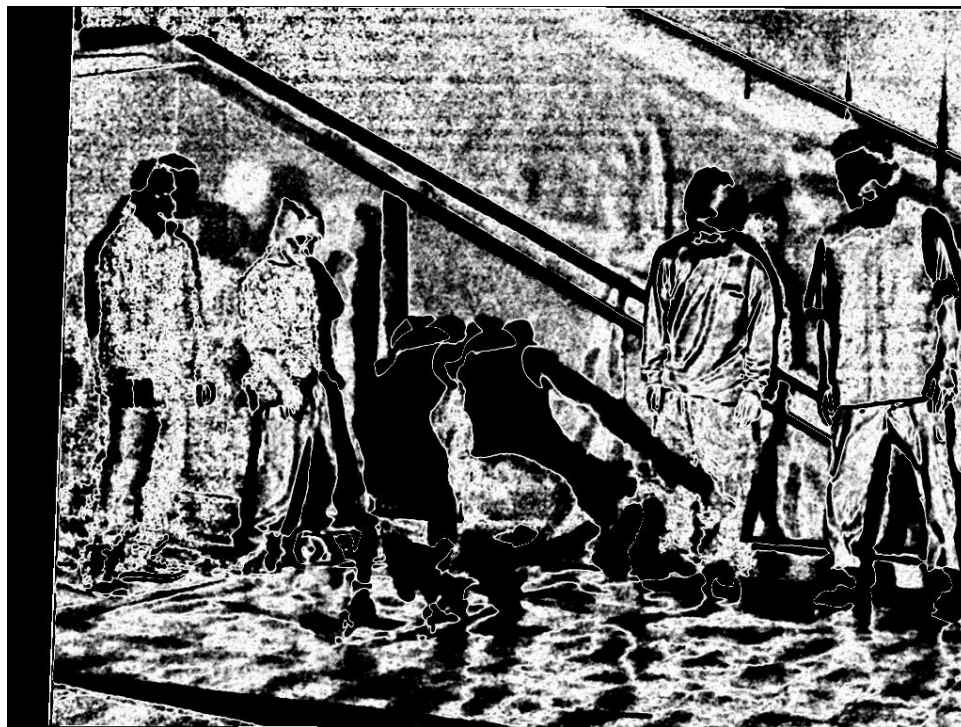
the verification map V^n.

The first particular solution we are interested in is the effect of using piecewise affine motion model based on depth segmentations instead of using a global motion model. Method a and c are not having layers and both are among the poorest performing methods. Figure 12-(a) also shows the registration weights I_W^0, which is significantly worse than the registration weights we obtained in any other method.

Our second investigation is on the effect of using sub-regions to further divide the depth segments. From the results and the observations on the registration of individual patches (see Figure 3) we can conclude that the estimation at the sub-regions has a positive effect. An optimal subdivision number R, however, cannot be decided since unexpected registration errors are always possible. It is definite on the other hand that increasing R number lengthening the total motion estimation time and the image patches become more likely to be registered incorrectly as their sizes decrease.

Method	L	R	Model	Initialization	Global+Subregion
a	1	1	affine	-	Subregion
b	11	1	translation	-	Global+Subregion
c	1	3	affine	-	Subregion
d	11	3	translation	-	Subregion
e	11	3	affine	calibration-init, subregion-init	Subregion
f	11	1	affine	-	Subregion
g	11	3	translation	-	Global+Subregion
h	11	3	affine	subregion-init	Subregion
i	11	3	affine	-	Subregion
k	11	3	affine	-	Global+Subregion
l	11	3	affine	subregion-init	Global+Subregion

Table 1 – Parameters for the different registration results

Figure 10 - Registration Performance
(using Registration Weights)

	a	b	c	d	e	f	g	h	j	k	l
Avg.	75.423	78.315	78.643	79.444	81.272	82.451	82.796	83.31	83.828	86.572	87.016

Figure 11 - Registration Performance
(after Reconstruction)

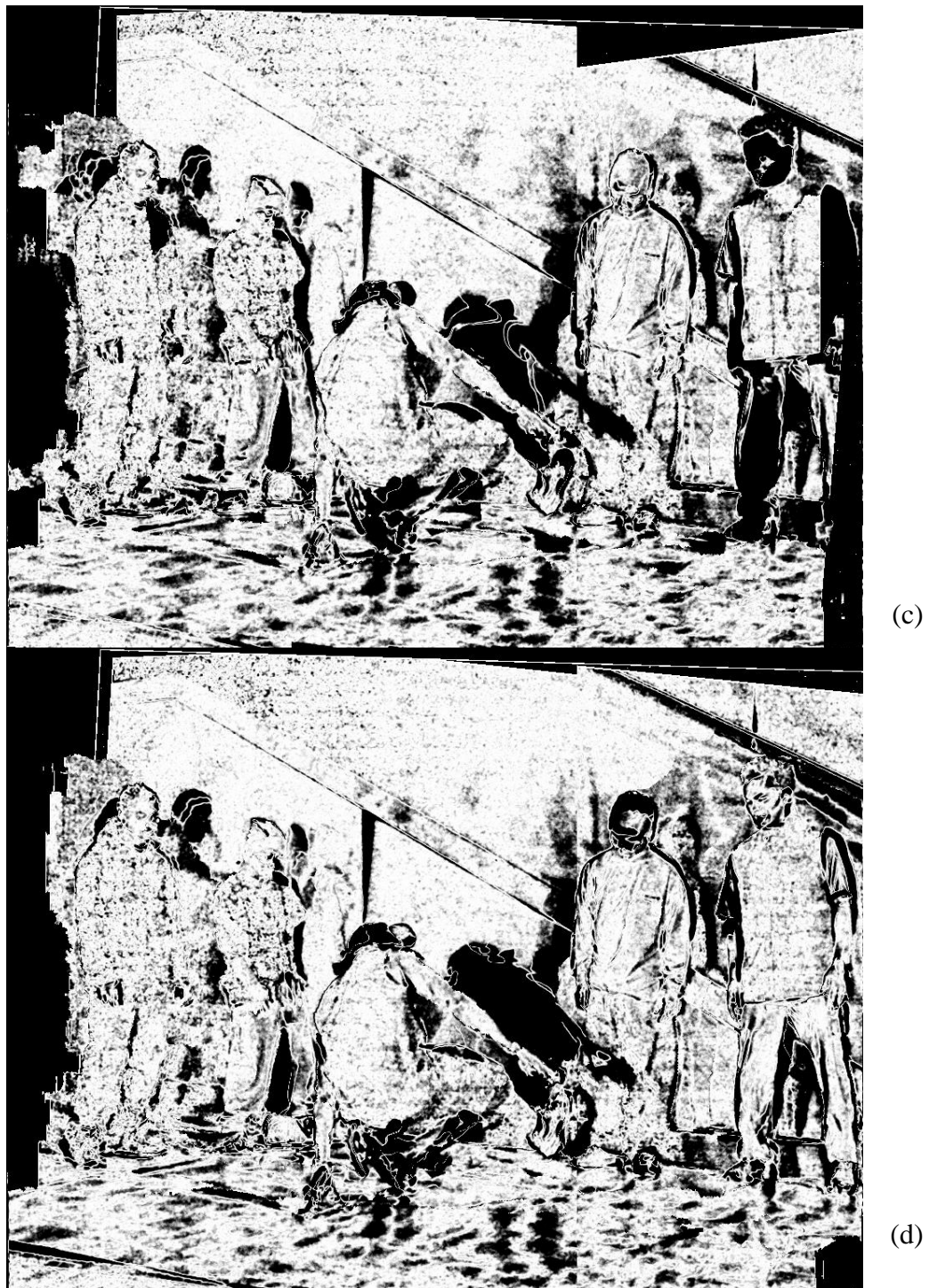|	a	b	c	d	e	f	g	h	j	k	l
Avg.	78.777	82.457	82.174	83.235	83.99	85.93	85.273	86.118	86.895	88.698	88.699

(a)

(b)

(c)

(d)

Figure 12 – Registration Weights: (a) for method a, (b) for method j, (c) for method k, (d) for method l.

5.3 Evaluation Results of Super-resolution

Two experiments are performed. In Set 1, LR images are generated artificially from the original HR views by filtering with a known h_{psf} and subsampling by the factor of 2. Super-resolved images are then generated using the known h_{psf}. In Set 2, images in the Breakdancing are directly used as LR inputs and h_{psf} parameters are estimated using the method stated previously. The best parameters are chosen to be $K_{psf} = 13$ and $\sigma_{psf} = 2.2$.

5.3.1 Quantitative Evaluation Results of Super-resolution

We use Peak Signal to Noise Ratio (PSNR) as the quantitative measure. We also provide 2D DFTs (Discrete Fourier Transforms) of the interpolated and super-resolved images to illustrate the improvement of the high-frequency content in the frequency domain.

The PSNR values for interpolated and super-resolved pairs, computed for Set 1 results, are presented in Table 2. It is observed that the PSNR of interpolated images are higher for both views, since artifacts and noise introduced by the multi-view super-resolution process negatively affects PSNR. In 3D viewing, however, artifacts in individual views become less observable and the overall enhancement in resolution is better perceived. We also note that super-resolved images are constructed from 8 views, and may contain some high-frequency components that are not contained in the original reference frame. These considerations motivate us to conduct 3D subjective evaluation of the resulting images.

	Left View	Right View
Bicubic Interpolation	37.897	38.867
Super-resolution	36.020	36.195

Table 2: Quantitative Evaluation Results for Set 1

A second option for the quantitative evaluation is to consider the results in the frequency domain. Since the aim of the super-resolution is to reconstruct high-frequency content, it is reasonable to inspect the results in frequency domain to assess the effectiveness of the super-resolution operation. Figure 13 shows the left view Set 1 results in the DFT domain. For the DFT result of the super-resolved image we see that the frequency contents are widened to higher frequencies those are missing for the interpolated ones. However, as in the PSNR results, the left and right views are treated separately and the effect of 3D perception is ignored in these quantitative results as well.
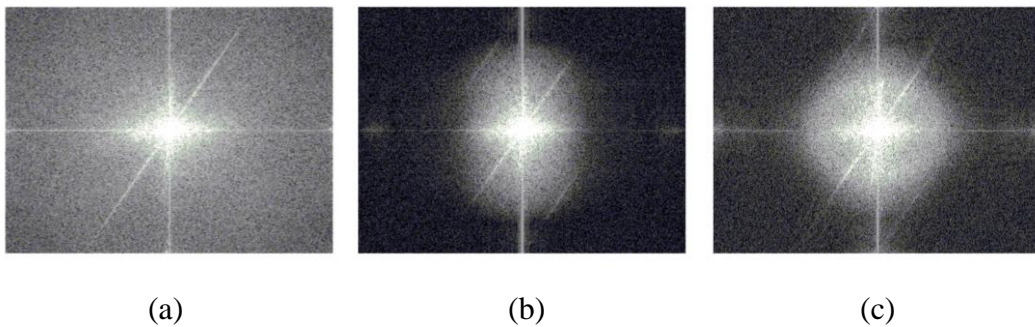

(a)   (b)   (c)

Figure 13 – 2D DFT Results for Set 1: (a) original, (b) bicubic interpolation, (c) SR.

5.3.2 Subjective Evaluation Results of Super-resolution

Portions from left-view SR images generated by three different image registration methods for Set 2 are shown in Figure 16. The grades given by the assessors are evaluated to compute overall test scores for each test image. Figures 14-15 and Tables 3-4 provide the test scores for each of the image sets. Lower test scores indicate a better perceived quality. The results are also depicted in line plots where the test scores are compared according to the 3D display types. The results show that SR images are consistently perceived better on both 3D displays.
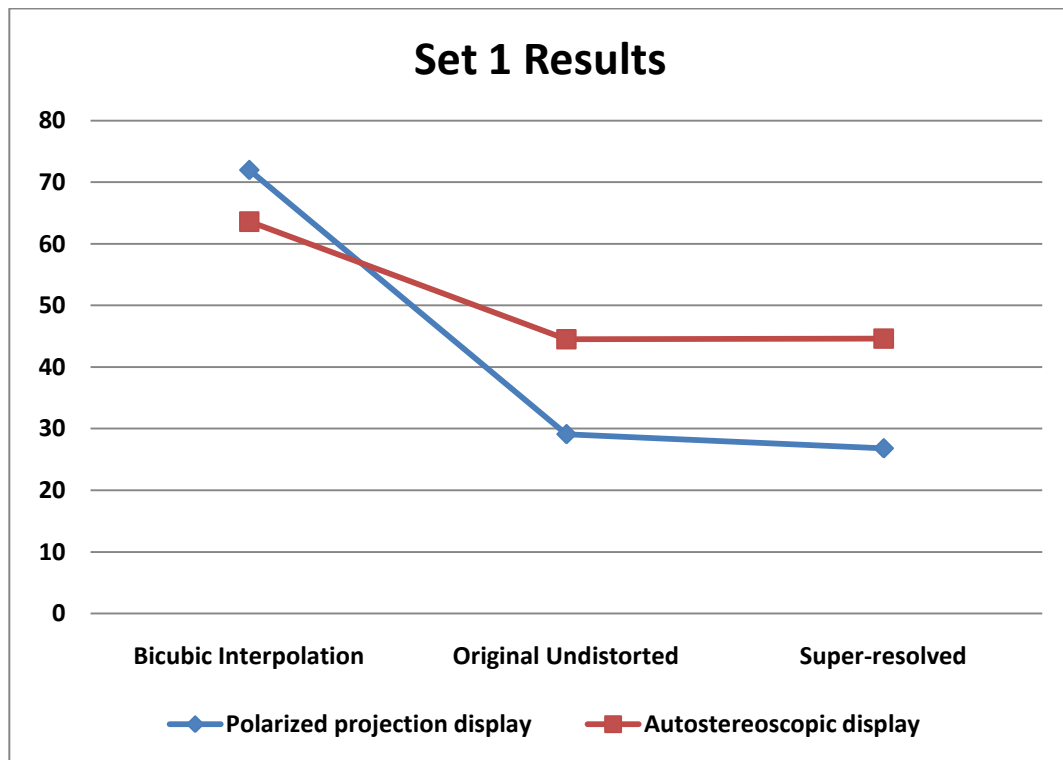
Figure 14 – Set 1 Subjective Evaluation Results

	Polarized projection display		Autostereoscopic display	
	Test Score	±		±
Bicubic Interpolation	72	2.8	63.6	1.5
Original Undistorted	29.1	5.2	44.5	9.7
Super-resolved	26.8	5.8	44.6	7.7

Table 3 – Set 1 Subjective Evaluation Results
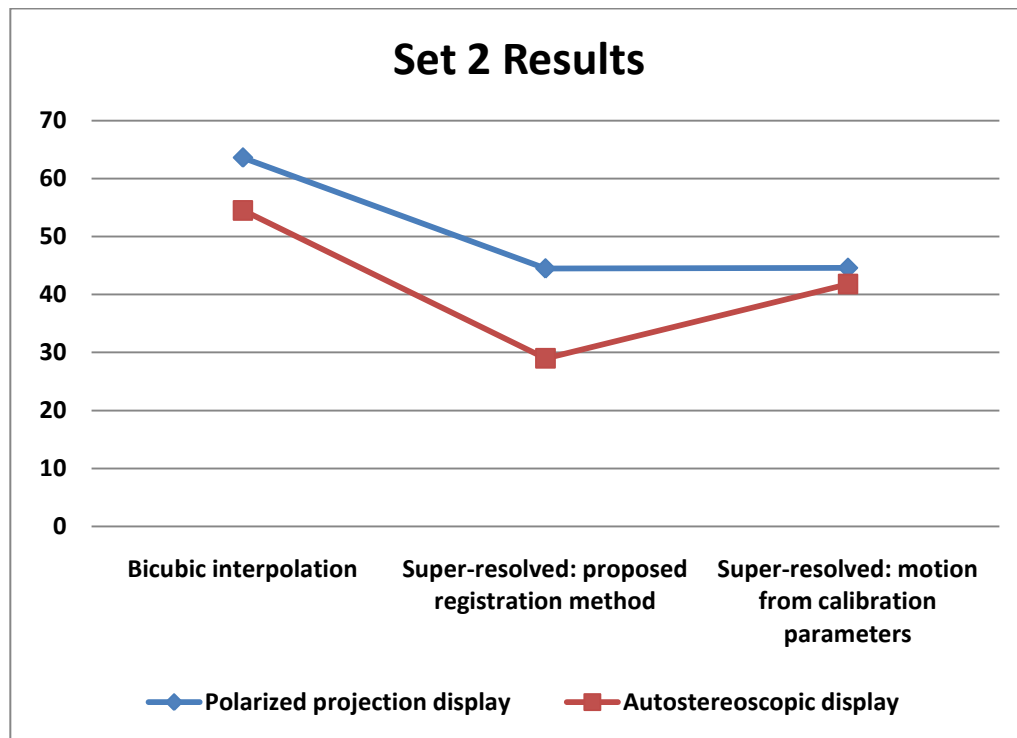
Figure 15 – Set 2 Subjective Evaluation Results

	Polarized projection display		Autostereoscopic display	
	Test Score	±		±
Bicubic interpolation	63.6	1.50	54.5	1.7
Super-resolved: proposed registration method	44.5	9.7	29	7.7
Super-resolved: motion from calibration parameters	44.6	7.7	41.8	7.2

Table 4 – Set 2 Subjective Evaluation Results

Figure 16 – Set 2 SR Results: (a) Bicubic interpolation (b) SR-proposed method without any parameter initialization, (c) SR-parameters initialized from calibration matrices (d) SR-by using motion from calibration only.

5.4 Effect of Wrong PSF Parameters on the SR Results:

We also investigated the effect of wrong PSF parameters and the sensitivity of the result to the accuracy of these parameters.

To test that effect, the 8 views are down-sampled by a factor of 2 with the known 2D Gaussian PSF with parameters of $K_{psf} = 7$ and $\sigma_{psf} = 1.7$ as in the previous case but this time the SR reconstruction is repeated using the PSFs with varying parameters.

Small portions of the results are presented in Figure 17 for varying σ_{psf} values and in Figure 18 for different K_{psf} values.

Our first observation is that a too small estimation of the kernel support K_{psf} has unacceptable effects on the resulting SR images. This result is reasonable since truncating the Gaussian PSF at its high valued tails deforms the characteristics of the Gaussian function.

Using too large K_{psf} has no observable effect on the result. This validates our strategy of firstly determining the correct σ_{psf} and then setting the K_{psf} to a large enough value that avoids truncation of the Gaussian function.

At the vicinity of the truncation, σ_{psf} can be chosen more flexibly at a reasonable range without causing the disturbing artifacts. The effect of wrong σ_{psf} is limited with a decrease in the SR performance and the results will have a reasonable amount of blur when compared with the SR results with the correct parameters.

Figure 17: Effect of σ_{psf} parameter on SR; (a) Bi-cubic interpolation, (b) σ_{psf}=1.1, (c) σ_{psf}=1.3, (d) σ_{psf}=1.7, (e) σ_{psf}=2.1, (f) σ_{psf}=2.3.

Figure 18: Effect of K_{psf} parameter on SR; (a) Bi-cubic interpolation, (b) K_{psf}=3, (c) K_{psf}=5, (d) K_{psf}=7, (e) K_{psf}=9, (f) K_{psf}=11.

Chapter 6

CONCLUSIONS

We propose a method for depth guided registration and super-resolution of multi-view video. We model the 3D scene as a collection of approximately planar image patches and registered these patches onto the reference HR grid using parametric models, which can provide sub-pixel accuracy. We tested our method on real data to generate super-resolved HR stereoscopic images. The results are evaluated by two methods: i) with ground truth (by first downsampling and then upsampling using the SR method), and ii) with no ground truth (by directly upsampling the actual images). The former is carried out by quantitative metrics and subjective tests, while the latter is done by subjective tests only. The results of subjective tests show successful SR of multi-view video has been achieved.

We see that piecewise parametric motion estimation along with the depth maps provide a suitable facility to be used in 3D SR and related applications such as view interpolation.

As a future application of the method, it is possible to construct a free viewpoint resolution enhanced 3D video sequence from a common linear multi-camera setup.

Although it is not adequately investigated in this work, super-resolution using a more common 2-view-stereo setup seems possible with the inclusion of the next and previous video frames in the sequence.

POCS algorithm performs well for those applications without a processing time requirement. For real-time applications any faster SR algorithm may directly replace the POCS algorithm in the method.

Piece-wise parametric motion models perform well in multi-view video even for a high accuracy demanding SR application. The multi-view image registration framework can have applications in other 3D video algorithms that require motion-compensation such as super-resolved view interpolation or inter-view 3D video coding.

The technique can handle occlusion problems, correspondence problem and aperture problem.

Resolution increase is definitely noticeable in 3D. The increase in the details and sharpness can easily be noticed and preferred by the human subjects. Although it is possible for any SR technique to introduce some degrading effects and artifacts due to registration errors (or the deficiency of the motion model), they are tend to be less perceived in 3D stereo images compared to 2D monocular images. This is due to the fact that the 3D perceived image composes of two images that are unlikely to share the same artifact and the artifacts are mostly a decrease in the details and can be easily compensated by the complementary view.

No calibration or prior information is needed for the method. When such additional information is available, they can be embedded into the proposed framework.

Handling boundaries is still a problem. The method does most of its enhancement at the approximately planar image regions. Although the depth images provide a cue for the object boundaries, since they are given in LR and not fully accurate using them directly may cause further artifacts.

Not all scene elements are suitable to be registered in high accuracy using the parametric motion models. Especially continuous edges along the depth direction can be problematic. Yet again, parametrically registered image patches can be an input of any refining motion estimation step.

Layer by layer and depth dependent processing is possible. Camera blurs depend on the distance of the subject from the camera. Using this technique, different PSF parameters can be used in the reconstruction of individual layers.

BIBLIOGRAPHY

[1] Irani, M. and Peleg, S., "Motion analysis for image enhancement: Resolution, occlusion, and transparency." Journal of Visual Communication and Image Representation, 1993, Vol. 4, pp. 1993-12.

[2] Tanaka, M., Yaguchi, Y. and Okutomi, M., "Robust and accurate estimation of multiple motions for whole-image super-resolution." 2008. pp. 649-652.

[3] Baker, S., Szeliski, R. and Anandan, P., "A layered approach to stereo reconstruction." 1998. INSTITUTE OF ELECTRICAL ENGINEERS INC (IEEE). pp. 434-441.

[4] Hsu, S. and Anandan, P., "Accurate Computation of Optical Flow by Using Layered Motion Represent at ions."

[5] Shade, J., et al., "Layered depth images." 1998. pp. 231-242.

[6] Zhang, Li, et al., "Multiple view image denoising." Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, s.l. : IEEE Computer Society, 2009, Vol. 0, pp. 1542-1549.

[7] Hardie, R.C., Barnard, K.J. and Armstrong, E.E., "Joint MAP registration and high-resolution image estimation using a sequence of undersampled images." IEEE Transactions on Image Processing, 1997, Vol. 6, p. 1621.

[8] Bishop, Christopher M., "Bayesian image super-resolution." s.l. : MIT Press, 2002. pp. 1303-1310.

[9] Capel, D. and Zisserman, A., "Computer vision applied to super resolution." IEEE Signal Processing Magazine, 2003, Vol. 20, pp. 75-86.

[10] Koch, R., Pollefeys, M. and Gool, L. Van., "Realistic surface reconstruction of 3D scenes from uncalibrated image sequences." The Journal of Visualization and Computer Animation, s.l. : John Wiley & Sons, 2000, Vol. 11, pp. 115-127.

[11] Baker, S. and Kanade, T., "Super-resolution optical flow." Robotics Institute, Carnegie Mellon Univ., Pittsburgh, PA, CMU-RI-TR-99--36, s.l. : Citeseer, 1999.

[12] Zhao, W.Y. and Sawhney, H., "Is super-resolution with optical flow feasible?" Computer Vision—ECCV 2002, 2002, pp. 599-613.

[13] Ince, S., et al., "Depth estimation for view synthesis in multiview video coding." 2007. pp. 1-4.

[14] Odobez, JM and Bouthemy, P., "Robust multiresolution estimation of parametric motion models." Journal of visual communication and image representation, s.l. : Elsevier, 1995, Vol. 6, pp. 348-365.

[15] Freeman, W.T., Jones, T.R. and Pasztor, E.C., "Example-based super-resolution." IEEE Computer Graphics and Applications, 2002, Vol. 22, pp. 56-65.

[16] Ur, H. and Gross, D., "Improved resolution from subpixel shifted pictures." CVGIP: Graphical Models and Image Processing, s.l. : Elsevier, 1992, Vol. 54, pp. 181-186.

[17] Schultz, R.R. and Stevenson, R.L., "Extraction of high-resolution frames from video sequences." IEEE Transactions on Image Processing, 1996, Vol. 5, pp. 996-1011.

[18] Stark, Henry and Oskoui, Peyma., "High-resolution image recovery from image-plane arrays, using convex projections." J. Opt. Soc. Am. A, s.l. : OSA, 1989, Vol. 6, pp. 1715-1726.

[19] Patti, A.J., et al., "Superresolution video reconstruction with arbitrary sampling lattices and nonzero aperture time." IEEE Transactions on Image Processing, s.l. : Citeseer, 1997, Vol. 6, pp. 1064-1076.

[20] Rec.BT.500-11, ITU-R., "Methodology for the subjective assessment of the quality of television pictures." *Methodology for the subjective assessment of the quality of television pictures.* 2002.

[21] Saygili, G., Gurler, G. and Tekalp, A.M., "3D display-dependent quality evaluation and rate allocation using scalable video coding." 2009. pp. 717-720.

[22] Zitnick, C.L., et al., "High-quality video view interpolation using a layered representation." 2004. ACM New York, NY, USA. pp. 600-608.

[23] Tekalp, A M., "The digital signal processing handbook." [ed.] DB Williams V. s.l. : CRC Press, 1998, Image and Video Restoration.

[24] Wiegand, T., et al., "Joint draft 11 of SVC amendment." Joint Video Team (JVT), Doc. JVT- X, 2007.

[25] Tekalp, A.M., Ozkan, M.K. and Sezan, M.I., "High-resolution image reconstruction from lower-resolution image sequences and space-varying image restoration." 1992. pp. 169-172.

[26] Patti, A.J. and Altunbasak, Y., "Artifact reduction for set theoretic super resolution image reconstruction with edge adaptive constraints and higher-order interpolants." IEEE Transactions on Image Processing, 2001, Vol. 10, p. 179.

[27] Park, S.C., Park, M.K. and Kang, M.G., "Super-resolution image reconstruction: a technical overview." IEEE signal processing magazine, 2003, Vol. 20, pp. 21-36.

[28] Irani, M. and Peleg, S., "Improving resolution by image registration." CVGIP: Graphical Models and Image Processing, 1991, Vol. 53, pp. 231-239.

[29] Eren, PE, Sezan, MI and Tekalp, AM., "Robust, object-based high-resolution image reconstruction fromlow-resolution video." IEEE Transactions on Image Processing, 1997, Vol. 6, pp. 1446-1451.

[30] Cheng, Xiaoyu, Sun, Lifeng and Yang, Shiqiang., "Generation of Layered Depth Images from Multi-View Video." 2007. Vol. 5, pp. V--225--V--228.

[31] Pentland, A.P., "Depth of scene from depth of field." s.l. : SRI INTERNATIONAL, 1982.

VITA

Özgün Genç was born in Ankara, Turkey in 1982. After completing his high school education in 60.Yıl Anadolu Lisesi, Izmir, he entered to Bilkent University, Ankara in 2000. He received the degree of Bachelor of Science in Electrical and Electronics Engineering from Bilkent University in 2004. Between 2006 and 2007 he worked as Software Design Engineer in Nortel-Netas Telecommunications, Istanbul. In 2007, he entered to the Graduate School of Sciences and Engineering of Koc University, Istanbul, and worked as a teaching assistant until 2009.