The Applications of Compressive Sensing in Multi-modal Images

Dissertation

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften

(Dr.-Ing.)

von

M.Sc. Juanjuan Han

eingereicht bei der Naturwissenschaftlich-Technischen Fakultät

der Universität Siegen

1. Gutachter: Professor Dr.-Ing.habil.Otmar Loffeld

2. Gutachter: Professor Dr.-Ing. Joachim Ender

Datum der mündlichen Prüfung: 25.08.2014

Abstract

In the past decade, an exciting new theorem, known as Compressive Sensing or Compressed Sensing (CS), mathematically establishes that relatively small number of non-adaptive, linear measurements can harvest all of the information necessary to faithfully reconstruct sparse or compressible signals. This leads to the reduction of sampling rates, storage volume, power consumption and computational complexity in signal and image processing.

The thesis develops three different applications of compressive sensing in multimodal images. The first application presents an effective multi-image fusion scheme based on a Discrete Cosine Transform (DCT) sampling model for compressive sensing imaging by taking advantage of the sparsity of the image in the spectral domain. In the second application, although the depth images delivered by 3D vision system based on Time-of-Flight (ToF) camera provide new perspective, they suffer from relatively low spatial resolution in comparison with color images due to the size limitation of current 2D pixel array in ToF sensor. Hence, the *soft* solution (i.e., post-processing) to increase the spatial resolution deserves to be advocated in comparison to high payoff of the *hard* solution (i.e., hardware improvement). From this point of view, the thesis attempts to explore the potential approaches within the framework of compressive sensing to enhance resolution of depth image. Regarding the third application, a rapid development of related research work with regard to processing and analysis of multi-modal image data is urgently desired due to the 2D/3D vision system which not only provides 2D view of the scene but also depth information of the same scene has become increasingly attractive. Hence, the thesis also attempts to explore potential approaches based on CS to sense change in the multi-modal images.

The work presented in the dissertation is expected to contribute to the related field by addressing the following aspects:

- For multi images fusion, in order to reduce the computational complexity and to save storage space, an effective fusion scheme based on compressive sensing is presented.

- To enhance the lateral resolution of depth image, a novel method is proposed by adopting the recent emerging theory of compressive sensing. This approach benefits from the finding of the sparsity of depth image, which is different from the conventional methods and open a new way for super-resolution reconstruction of depth image.

- In the point of view of the sparsity in image analysis, the thesis presents an innovative approach to detect change occurred in multi-modal images. The proposed approach mainly focuses on sparse feature pursuit with arbitrary shape and reconstruct via matrix decomposition. So far, to our knowledge, matrix decomposition has not yet been applied to the multimodal image data. Thus, this formulation yields a novel model for this application.

Kurzfassung

Die Forschungstätigkeitenin „Compressive Sensing“ (auch „Compressed Sensing“) haben während der letzten Dekade zu sehr interessanten und aufregenden Ergebnissen im Bereich der Signalverarbeitung beigetragen.Hierbei kann mathematisch belegt werden, daß unter gewissen, aber definierten Umständen, durch einige wenige, nicht-adaptive und lineare Messungen eine vollständige Wiederherstellung kompressibler Signale möglich sein kann. Die Anwendung dieser Technik ermöglicht dann, um Beispiele anzuführen,die Verringerung der Signalabtastraten,der Speichervolumina zur Sicherung von Daten sowie der Komplexität in Bild- und Signalverarbeitung.

In dieser Arbeit werden drei verschiedene Anwendungen dieses Forschungsgebiets entwickelt und präsentiert.

Die erste Anwendung zeigt eine Lösung für eine effektive Methode zur Fusionierung von Daten auf, die aus Multi-Sensoren-Systemen stammen(ToF-Kameras), die auf einer diskreten Kosinus-Transformation (DCT) basiert.Hierbei wird davon profitiert, dass nach der Transformation die Spektraldarstellung der Daten in der neuen Basis „sparse“ ist.

Die zweite Anwendung konzentriert sich auf die Verbesserung von Tiefenbild-Aufnamen der ToF-Systemen, die durch die Möglichkeit der Aquisitionvon 3D-Distanz-Daten eine neue Perspektive in die Bewertung solcher Datensätze einbringen.Diese Abstandsdatensind üblicherweise von geringererräumlicher Auflösung als im Vergleich zu herkömmlichen 2D-Darstellungen. In einem nachgelagerten Schritt kann diesem Nachteil durch „Hochrechnen“ der Bildauflöung entgegengewirkt werden („soft solution“). Alternativ ist es durch Veränderung der Hardwareausstattung möglich, einen ähnlichen Effekt zu erzielen (e.g. Austausch der Sensoren mit höherer Auflösung).Im Rahmen des „Compressive Sensing“ werden hierbei Wege und Möglichkeiten untersucht, den kostenintensiven Austausch der Hardware zu vermeiden.

Durch die inzwischen kostengünstige und einfache Verfügbarkeit von ToF-Systemen istauch der Wunsch entstanden nach einfachen Möglichkeiten, in Bild- und

Videoszenen Änderungen (e.g. Objektbewegungen) festzustellen. Eine dritte und die damit letzte Anwendung, die in dieser Dissertation untersucht wird,ist damit, wiesolche Änderungen in multi-modalen Datensätzenmithilfe des„Compressive Sensing"-Frameworksdetektiert werden können.

Zusammenfassend kann der Beitrag dieser Arbeit damit auf die folgenden Aspekte festgelegt werden:

- Eine Methode, um Daten (Bilder) unterschiedlicher Sensoren einheitlich durch „Compressive Sensing" behandeln zu können, z.B., um Speicherplatz zu sparen oder die Komplexität in Berechnung zu reduzieren.

- Erhöhung der Bildauflösung in Tiefenbildaufnahmen (z.B. von ToF-Systemen, wie der MultiCam des ZESS). Dazu wurde eine neue Methode entwickelt, die eine der Grundannahmen des „Compressive Sensings" ausnutzt („sparsity"). Diese neuartigeHerangehensweise eröffnet eine Möglichkeit zur Steigerung der lateralen Auflösung („Superresolution").

- Mithilfe der „sparsity"-Eigenschaft wird eine neuartige Methode präsentiert, um Bewegungen in multi-modalen Bildaufnahmen zu detektieren und verfolgen. Die Rekonstruktion dieser Eigenschaften geschieht durch die Formulierung (und numerischer Lösung) eines zugeordneten Optimierungsproblems unter zuhilfenahme der „Lagrange"-Multiplikatoren.Die betreffenden Matrizen beschreiben dabei den Hinter- bzw. Vordergrund der Bildszene. So weit wir wissen, wurde dieses Verfahren noch nicht auf multi-modale Daten angewendet.

Acknowledgments

First of all, I would like to take this opportunity to express my deepest gratitude to my supervisor, Prof. Otmar Loffeld for his constant encouragement and invaluable guidance during my Ph.D study. I would like to express my sincere appreciation to Dr. Klaus Hartmann, the team leader of our research group, for his excellent advice and guidance regarding my research work.

I would like to give many thanks to those who helped me at the Centre of Sensor Systems (ZESS) in different way: Mussab Zubair, Stefan Lammars, Seyed Eghbal Ghobadi, Omar Edmond Loepprich, Oliver Lottner, Benjamin Langmann, Wolfgang Weihs, Sven Stark, Wolf Twelsiek, Rolf Wurmbach, Renate Szabo, Silvia Niet-Wunram, Katharina Haut.

I also would like to give my special thanks to my family, especially to my loving husband Robert Yu Wang for his continuous supporting and encouraging. I am deeply appreciated by inspiration from my parent and my parent in-law.

Last but not least, I would like to thank God for giving me this wonderful life!

Contents

List of Abbreviations

2D	Two dimensional
3D	Three dimensional
ToF	Time-of-Flight
CMOS	Complementary Metal Oxide Semiconductor
PMD	Photonic Mixer Devices
CS	Compressive Sensing
RIP	Restricted Isometry Property
NP	Non-deterministic Polynomial-time
SVD	Singular Value Decomposition
ALM	Augmented Lagrange Multiplier
DCT	Discrete Cosine Transformation
LASSO	Absolute Shrinkage and Selection Operator
MP	Matching Pursuit
OMP	Orthogonal Matching Pursuit
ROMP	Regularized Orthogonal Matching Pursuit
StOMP	Stagewise Orthogonal Matching Pursuit
CoSaMP	Compressive Sampling Matching Pursuit
BP	Basic Pursuit
LP	Linear Programming
LARS	Least Angle Regression
GPSR	Gradient Projection for Sparse Reconstruction
IST	Iterative Splitting and Thresholding
FPC	Fixed Point Continuation
PCAT	Principal Component Analysis Transform
RPCA	Robust Principal Component Analysis
ALM	Augmented Lagrange Multiplier Method
SVD	Singular Value Decomposition
WPT	Wavelet Packet Transform
FFT	Fast Fourier Transformation

CT	Computed Tomography
MRI	Magnetic Resonance Image
IE	Image Entropy
MI	Mutual Information
AG	Average Gradient
TDC	Time-to-Digital Converter
SR	Super-Resolution
HR	High-Resolution
LR	Low-Resolution
MAP	Maximum-A-Posterior
MRF	Markov Random Fields
JBU	Joint Bilateral Upsampling
JBF	Joint Bilateral Filter
JPEG	Joint Photographic Experts Group
MPEG	Moving Picture Experts Group
FPGA	Field Programmable Gate Array
PSNR	Peak Signal-to-Noise Ratio
VGA	Video Graphics Array

List of Figures

List of Tables

1 Introduction

On one hand, from the point of view of compressive sensing, the research work presented in the thesis is an attempt to explore the potential approaches to solutions for multi-image fusion and super-resolution reconstruction of depth image, respectively; and on the other hand, from the perspective of low-rank and sparse matrix recovery, the thesis also attempts to explore the approach to detecting change in multimodal image provided by a 2D/3D vision system.

This is an introductory chapter that outlines the motivation, objective and contribution of the thesis.

1.1 Motivation

With the rapid development of sensor systems, the information science focuses mainly on how the information about the real world is extracted from the sensor data. In many cases, a single sensor is not sufficient to provide a complete and fully informative perception of the real world. Therefore, multi-sensor fusion has attracted a great deal of attention in the past years. Image fusion is a branch of multi-sensor fusion and refers to a process of combining relevant information from two or more images into a fused image that possesses more information than any of the input images. The current image fusion schemes can be classified roughly into pixel-based and region based methods. For both of them all the samples of the images have to be acquired, which means that the storage burden and the processing challenges must be handled especially due to the growing sensor data volumes. Recently, an exciting new field, Compressive Sensing (CS), also called compressed sensing or compressive sampling, has attracted considerable attention in areas of applied mathematics, computer science, and electrical engineering by suggesting that it may be possible to surpass the traditional limits of sampling theory. The CS theory exploits the knowledge that the signal or image we are acquiring is sparse in some known transform domain, which means that the signal or image is compressive. Then the compressive signal may be reconstructed accurately with sub-Nyquist data sampling rate from a significantly smaller number of measurements than sampling the original signal at Nyquist-Shannon rate. This is a clear and striking advantage compared with the conventional signal theory based on the Nyquist-Shannon sampling theory. The CS theory can lead to the reduction of sampling rates, storage volume, power consumption, and computational complexity in signal and image processing and related research fields. Based on the unique advantages that CS theory framework possesses, the first topic studied in the thesis is therefore orientated to explore the solution for multi-image fusion by using the CS theory framework.

As a recent development in imaging hardware, the Three-Dimensional (3D) Time-of-Flight (ToF) cameras has been introduced that use active sensing to capture 3D range/depth data at frame-rate as a per-pixel depth. A light source from the camera emits a near-infrared wave which is then reflected by the scene and is captured by a dedicated sensor. Depending on the distance of the objects in the scene, the captured light wave is delayed in phase compared to the original emitted light wave. By measuring the phase delay, the distance between the object in a scene and the camera at each pixel can be estimated. However, the depth map of the current ToF sensors suffers from the limitation in lateral resolution due to the restriction of the range sensor and therefore the weakness makes such kind of sensors inefficient for some applications in which the High-Resolution (HR) image data is required. Besides *hard-method* which increases the size of sensor array, most approaches pay more attention to *soft-method* which is from perspective of algorithms by means of image processing. Super-Resolution (SR) is a class of techniques that enhance resolution of image and is known as an ill-posed inverse problem. In recent years, the novel theory of CS emerged and paved a way to solve the ill-posed inverse problem. Accordingly, the second topic studied is orientated to explore an approach to SR problem of depth image by putting the SR problem of the depth image into the CS theory framework.

Detecting region of change in images of the same scene taken at different time is of widespread interest due to a large number of applications in the diverse disciplines. Important applications of change detection include video surveillance, remote sensing, civil infrastructure, and so forth. In spite of the diversity of applications, most work has mainly concentrated on the vision system that only operates a visible spectrum camera (e.g. 2D color images) and ignores the other sensor modalities to some extent. The combinational utilization of ToF camera and standard color camera has emerged recently as an unusual potential to spread due to the fact that on one hand it provides color image and on the other hand it delivers depth information between the camera and the object in a scene. In recent years, a new monocular 2D/3D imaging system called MultiCam [5] has been developed in our research center. Since change detection in multimodal image has been becoming more and more attractive for many applications, and in the meantime, the mathematic theory in low-rank and sparse matrix recovery obtained more attentions from researchers and developed dramatically, in view of this, detecting change in multi-modal images via way of matrix decomposition and recovery is proposed as the third topic of the thesis.

In a word, in this thesis, the topics studied respectively are multi image fusion via compressive sensing, super-resolution of depth image and change detection in multimodal image.

1.2 Objective of the thesis

The general objective of this thesis is threefold:

- The first objective is to use the newly emerged theory of CS to implement multi image fusion.
- The second one is to improve the quality of 3D depth image, especially lateral resolution which is the key aspect of depth image.
- The third one is to investigate an active and of widespread topic in many applications, i.e., change detection in multimodal images.

The methodological goal of the thesis is summarized as the following aspects:

- To explore a promising method based on the theory of compressive sensing to implement multi image fusion.

- To provide a promising method based on the basis of framework of compressive sensing to enhance the lateral resolution of depth image.

- To develop a promising method that can go far beyond the fact limit that depth image can only play an assistant role to 2D color image in multimodal image analysis. It can give exactly estimation of change region instead of the simple means via threshold technique. To this end, the thesis presents a sparse feature pursuit based approach to exactly reconstruct the region of change in multimodal image.

1.3 Thesis contributions and outline

The dissertation contributes to the area which is related to multimodal image analysis and processing. The main points of contributions can be summarized as the following aspects:

- To implement the multi image fusion, an approach based on the recent emerging theory of compressive sensing is proposed. Compared with the traditional approaches such as pixel-based and region-based methods which need to acquire all the samples of the images and thus results in storage burden and processing challenges, the proposed method on one hand leads to the reduction of sampling rate, storage volume, and power consumption; On the other hand it opens a door to fuse multi image by using the newly emerged theory of compressive sensing instead of conventional sampling theory.

- To enhance the lateral resolution of depth image, a novel method is proposed by adopting the theory of compressive sensing. This approach benefits from the finding of the sparsity of depth image and open a new way for super-resolution reconstruction of depth image.

- In the point of view of the sparsity in image analysis, the thesis introduces an innovative approach to detect change occurred in multi-modal images simultaneously provided by 2D/3D vision system in the same scene. The proposed approach mainly focuses on sparse feature pursuit with arbitrary shape and reconstruct via matrix decomposition. So far, to our knowledge, matrix decomposition has not yet been applied to the multimodal image data. Thus, this formulation yields a novel model for this application.

The thesis is structured into six chapters. We start with an introductary chapter which includes motivation, objective and key contribution. In Chapter 2, we give a background review of the compressive sensing theory and its relevant extension. Multi image fusion via compressive sensing will be presented in Chapter 3. In Chapter 4 super-resolution restruction of depth image in a sparse way will be studied. Detecting moving object in multimodal image using a sparse way will be presented in Chapter 5. And finally Chapter 6 summarizes this thesis and presents disscussions and outlook concerning some points.

2 Background

2.1 Compressive sensing

For many years, traditional signal processing has relied on the Nyquist-Shannon sampling thereom, which states that the number of samples required to capture a signal must be determined by the signal's bandwidth [8][9]. That is, to reconstruct a signal of band-limit, the traditional signal processing approaches sample the signal uniformly at a rate which is at least twice the bandwidth of the underlying signal. The methods need large storage space to save the measurements, and most often result in a waste of resources since the measurements are simply discarded after signal compression in many practical applications. An alternative sampling theory, well-known compressive sensing, turns the Nyquist-Shannon theory on its head and has spurred resurgence in the field of sparse signal processing with contributions from the applied mathematic, geometric functional analysis, electrical engineering and the theoretical computer science communities. The key idea behind compressive sensing is to accurately acquire signals from relatively few samples. CS opens up an innovated framework to jointly measure and compress signals that allows less sampling and storage resources than the traditional approaches based on Nyquist-Shannon sampling.

2.1.1 Sparse signal

The CS theory builds upon the assumption that the signal is sparse under some basis. Before getting into the formulation of compressive sensing theory framework, some terminologies are first defined as follows:

- N-dimensional signal

  The dimension of a signal is the indepent components in a signal. Thus, "a vector signal is an N-dimensional signal" means a signal is an $N \times 1$ vector signal.

- K-sparse

  K is the sparsity number of a signal. “An N-dimensional signal f is K-sparse“ means that among the N components of f, only K of them are non-zero and $N-K$ are zero.

- An orthonormal basis

  An orthonormal basis for an inner product space V with finite dimension is a basis for V whose vectors are orthonormal.

Consider a real-world, finite-length, one-dimensional, discrete time signal f, which can be represented as a $N\times 1$ column vector in an N-dimensional space $\mathbb{R}^N$ with elements $f[n], n=1,2,...,N$. Any signal f in $\mathbb{R}^N$ can be represented with a basis of $N\times 1$ vector $\{\psi_i\}_{i=1}^N$ as

$$f=\sum_{i=1}^{N} x_i\psi_i \quad . \tag{2.1}$$

If the $N\times N$ matrix Ψ is used as basis matrix and assumed to be orthonormal, the signal f also can be represented in matrix form, as shown below:

$$f=\Psi x \quad , \tag{2.2}$$

where x denotes the coefficient sequence of f using the basis $\{\psi_i\}_{i=1}^N$. It is obvious that x and f are equivalent representations of the same signal with x in time domain and f in a certain transform domain Ψ.

2.1.2 Compressive measurements

In the framework of compressive sensing, measurements are taken not by directly sampling the sparse signal but by measuring a few of linear projections of the underlying signal. The linear measurements can be modeled as

$$y = \Phi f \, , \tag{2.3}$$

where $y = \left[y_1, y_2, ..., y_M \right]^T$ denotes the measurements, Φ is an $M \times N$ projection matrix, and the number of measurements M is far less than the signal dimension N. Let $\Phi = \left[\hat{\phi}_1, \ \hat{\phi}_2, \ \ldots, \hat{\phi}_M \right]^T$, then one measurement is supposed to project the signal f onto $\hat{\phi}_1$. The inner product can be measured as

$$y_i = \left\langle f, \hat{\phi}_i \right\rangle \, , \tag{2.4}$$

where $i \in \{1, 2, ..., M\}$. A reduced sampling rate is achieved if the inner product is made in the analog domain.

With the priori information that f is sparse in some basis Ψ, given the linear measurements y, the sparse signal f can be exactly reconstructed from y via nonlinear optimization.

However, the inverse problem is a highly underdetermined problem in general since $M \ll N$. To recover the signal with high probability, the projection matrix (also called sensing matrix or measurement matrix) must fulfill some properties. A crucial factor is the incoherence between the sensing matrix Φ and the sparsity basis Ψ. The incoherence means that the orthogonal projection will spread out information of sparse (highly localized) signals in the entire projection space and thus makes them insensitive to "under-sampling". That is, $\hat{\phi}_i$ cannot be represented on Ψ, and vice versa [1][9][11].

Incoherence:

The coherence between two matrices is defined by:

$$\mu(\Theta)=\mu(\Phi, \ \Psi)=\sqrt{N}\max_{1\le k,j\le N}\frac{\left|\left\langle \phi_k,\psi_j\right\rangle\right|}{\left\|\phi_k\right\|_2\left\|\psi_j\right\|_2}, \tag{2.5}$$

where ϕ_k is the k^{th} row of the sensing matrix Φ, ψ_j is the j^{th} row of the orthogonal basis Ψ and $1\le\mu\le\sqrt{N}$. The incoherence property requires that the rows $\{\phi_k\}$ of Φ cannot sparsely represent the columns $\{\psi_j\}$ of Ψ, and vice versa. A small μ means the sparse mapping operator will spread out information of sparse coefficients over the entire measurement space and therefore make them insensitive to random under-sampling, otherwise the reconstruction of non-zero coefficients will be biased towards certain positions. Compressive sensing is mainly concerned with low coherence pairs [2]. The examples of such kind of pairs are:

<u>The identity matrix and Fourier basis</u>

Since the fact that Φ is the sensing matrix corresponding to the classical sampling scheme in the time or space domain. The time-frequency pair obeys $\mu=1$ and therefore has maximal incoherence. Further, spikes and sinusoids are maximally incoherent not only just in one dimension but also in any number of dimensions. If it is the identity matrix, coefficients are identical to the signal. This allows us to employ the numerous powerful sparse reconstruction techniques for spectral estimation when the signal itself is sparse in an identity matrix basis in the context of CS [109].

<u>The noiselet matrix and Wavelet basis</u>

The coherence between noiselets and Haar wavelet is $\sqrt{2}$ and that between noiselets and Daubechies D4 and D8 wavelets is, respectively, about 2.2 and 2.9 across a wide range of sample sizes N [2]. Noiselets are also maximally incoherent with spikes and incoherent with the Fourier basis. Since most image data are sparse under a wavelet basis, there exists great potential for using the framework of compressive sensing theory for image processing [110].

Random matrix and any fixed basis

The random matrices are the matrices consisting of random vectors that have a flat power spectral density and they are incoherent with any fixed basis with coherence of about $\sqrt{2\log n}$. The most known examples of such kind of matrices are random waveform whose entries are samples of independent and identically distributed (i.i.d.) random variables from Gaussian or Bernoulli/Rademacher (random ± 1) distributions [83].

The incoherence is also related to equivalent property, which is associated with $\Theta(\Theta = \Phi\Psi)$, called Restricted Isomety Property (RIP) that generalizes the notion of incoherence. The incoherence and the RIP are the bargaining chip to find the unique sparse solution.

Restricted Isometry Property:

The restricted isometry property is a concept that was introduced by Candes and Tao [12] and has been proved to be very useful in studying the general robustness of CS. The RIP provides a tool for determining sufficient conditions that guarantees the sparse recovery in the presence of noise. The conditions derived based on the RIP are deterministic [12], in other words, there is no possibility of failure.

Assume that $\Theta = \Phi\Psi$, Θ meets the RIP of order s if there exists a constant $\delta \in (0,1)$ for which

$$(1-\delta_s)\|\upsilon\|_2 \le \|\Theta\upsilon\|_2 \le (1-\delta_s)\|\upsilon\|_2 \tag{2.6}$$

holds for all s-sparse $\upsilon \in \mathbb{R}^N$. The smaller δ_s is, the better the sparse signal can be recovered in the presence of noise. As a matter of fact, the RIP presents that a sensing matrix will be valid if every possible set of υ columns of Θ forms an approximate orthogonal set and therefore preserves the energy of all vectors having only non-zero elements at the same K positions. The examples of matrices that have been proven to satisfy the RIP include independent and identically distributed Gaussian random matrices, Bernoulli matrices, and partial Fourier matrices [13].

As mentioned before, with these conditions including sparsity of signal, and incoherence or RIP in hand, to make reconstruction of original signal possible, an optimization technique must be used, and this is the topic of the next sub-section.

2.1.3 Sparse signal reconstruction

The reconstruction algorithms often rely on an optimization, which searches for the sparsest coefficients x_0 that agree with the measurements y. If M is sufficiently large and x_0 is strictly sparse, x_0 is the solution to the l_0 minimization:

$$\hat{x}_0 = \arg\min \|x\|_0 \quad s.t. \quad y = \Phi\Psi x \quad (2.7)$$

However, to solve this l_0 minimization is Non-deterministic Polynomial-time hard (NP-hard) problem [14]. Fortunately, the revelation that supports the CS theory is that a computationally tractable optimization problem yields an equivalent solution. We need to replace the l_0 minimization with l_1 minimization:

$$\hat{x}_1 = \arg\min \|x\|_1 \quad s.t. \quad y = \Phi\Psi x \quad (2.8)$$

The l_1 optimization problem, also known as basis pursuit [15], can be solved by linear programming approaches. However, the l_1 optimization problem requires cubic computation in general and therefore the cubic complexity renders it impractical for many applications. For this reason, a flurry of research on faster algorithms has been motivated and the work has been done to find alternative algorithms that are faster or give superior reconstruction performance.

In practice, the linear projection measurements are contaminated with noise and the measurements can be modeled by

$$y = \Phi\Psi x + \varepsilon, \quad (2.9)$$

where ε denotes zero-mean white Gaussian noise. Then the recovery algorithm must consider the effect of noise. Basic Pursuit Denoising (BPDN) was proposed to solve such kind of model [3]:

$$\min \|\Phi\Psi x - y\|_2^2 + \lambda \|x\|_1 \quad s.t. \quad y = \Phi\Psi x + \varepsilon \, , \tag{2.10}$$

where $\lambda\left(\lambda > 0\right)$ is the balance parameter that balances the tasks of minimizing the l_2 norm of the noise and the minimization of the l_1 norm of the sparse signal.

The l_1 norm minimization with the presence of noise can also be formulated as a Least Absolute Shrinkage and Selection Operator (LASSO) problem [1]:

$$\min \|x\|_1 \quad s.t. \quad \|\Phi\Psi x - y\| \le \eta \, , \tag{2.11}$$

where η limits the noise power in the measurements.

The incomplete collection of the existing algorithms for reconstruction of the signal from the measured signal is listed as below:

- Iterative greedy algorithms such as Matching Pursuit (MP) [84] and its popular extensions such as Orthogonal Matching Pursuit (OMP) [18], Regularized Orthogonal Matching Pursuit (ROMP) [16], Stagewise Orthogonal Matching Pursuit (StOMP) [30], LASSO [85] and Compressive Sampling Matching Pursuit (CoSaMP) [17].

- Algorithms based on convex optimization methods such as Basis Pursuit (BP) [86], Linear Programming (LP) decoding [12], Least Angle Regression (LARS) [88] and Gradient Projection for Sparse Reconstruction (GPSR) [89].

- Iterative thresholding such as, Iterative Splitting and Thresholding (IST) [90], Bregman iterative algorithm and Fixed Point Continuation (FPC) [92][91] and a successor of FPC called FPC_AS [93].

An incomplete collection of various sparse reconstruction toolboxes have been available for solving l_1 norm minimization problem as below:

- **L1-MAGIC**

Contributions: L1-MAGIC is a collection of MATLAB routines for solving the convex optimization programs central to compressive sampling. The algorithms are based on standard interior-point methods, and are suitable for large-scale problems.

Download link: http://users.ece.gatech.edu/~justin/l1magic/

- **SparseLab**

Contributions: SparseLab is a library of Matlab routines for finding sparse solutions to underdetermined systems.

Download link: http://sparselab.stanford.edu/

- **l1_ls**

Contributions: Matlab implement of the interior-point method for l_1-regularized least squares and solves an optimization problem of the form: $\min \|Ax-y\|_2^2 + \lambda \|x\|_1$.

Download link: http://www.stanford.edu/~boyd/l1_ls/

- **GPSR**

Contributions: a Matlab solver of gradient projection type for convex quadratic program.

Download link: http://www.lx.it.pt/~mtf/GPSR/

- **SPGL1**

Contributions: a Matlab solver for large-scale one-norm regularized least squares.

Download link: http://www.cs.ubc.ca/~mpf/spgl1/

- **YALL1 Group**

Contributions: a MATLAB package for various group/joint sparse recovery problems.

Download link: http://yall1.blogs.rice.edu/

- **YALL1**

Contributions: a MATLAB package for various l_1-minimization problems, using a dual alternating direction method.

Download link: http://yall1.blogs.rice.edu/

- **cvx**

Contributions: Matlab software for disciplined convex programming.

Download link: http://cvxr.com/cvx/

- **FPC**

Contributions: l_1-miminization using iterative shrinkage and continuation.

Download link: http://www.caam.rice.edu/~optimization/L1/fpc/

- **FPC_AS (Fixed-point continuation and active set)**

Contributions: an active-set acceleration of FPC.

Download link: http://www.caam.rice.edu/ optimization/L1/FPC_AS/

2.2 Low-rank and sparse matrix decomposition

Matrix representation of complex systems and models attracting more attentions in various areas often have the character that such a matrix consists of a sparse component and low-rank component. Practically, it is significantly interesting to take advantage of the decomposable character of such a complex system. More recently, the extension of CS technique to the recovery of low rank and sparse matrix has become a focus of research and is demonstrating a rapidly growing array of important applications. In some cases this leads to underlying matrix-based signal model with sparsity and low rank. In this section, we briefly examine problem of low-rank and sparse matrix decomposition from a theoretical perspective.

The fundamental mathematic problem is considered that the observation matrix I is represented as the sum of an unknown sparse matrix F and an unknown low-rank matrix B, which is described as $I = B + F$. To exactly recover the two components, it is more intuitive to consider applying l_0-norm (i.e., the number of non-zero entries) to control the sparsity structure in the matrix and matrix rank to encourage the low-rank structure, that is:

$$\min_{B,F} rank\left(B\right) + \lambda \left\|F\right\|_0 , \tag{2.12}$$

where $rank(\cdot)$ is the rank of matrix, $\|\cdot\|_0$ denotes the l_0-norm of matrix, λ is the non-negative balance parameter that trades off the rank of matrix B versus the sparsity of matrix F. However, the minimization is not directly tractable due to the fact that the major difficulty on one hand lies in the non-convexity of $rank\left(B\right)$ and on the other hand is that it is extremely difficult to minimize the function of l_0-norm. Hence, the decomposition problem of Eq. (2.12) is NP-hard in general and there is no effective solution to it. However, a computationally tractable alternative which is recently well-studied, that is, the convex relaxation is considered to be firstly performed on it.

Let the function $f : \mathbb{C} \to \mathbb{R}$, where $\mathbb{C} \subseteq \mathbb{R}^{d \times m}$. The convex hull [20] of f on $\mathbb{C}$ is defined as the largest convex function g so that $g\left(x\right) \le f\left(x\right)$ for all $x \in \mathbb{C}$. The nuclear norm or the trace norm $\|\cdot\|_*$ has been known as the convex hull of the $rank(\cdot)$ [21]:

$$\left\|B\right\|_* \le rank\left(B\right), \quad \forall B \in \mathbb{C} = \left\{B \left\| B \right\|_2 \le 1\right\}. \tag{2.13}$$

And the l_1-norm is the convex envelope of the l_0-norm [20]:

$$\left\|F\right\|_1 \le \left\|F\right\|_0 , \forall F \in \mathbb{C} = \left\{F \left\|F\right\|_\infty \le 1\right\}. \tag{2.14}$$

Both of the nuclear norm and the l_1-norm functions are convex but non-smooth, and they have exhibited to be effective surrogates of the matrix rank and of the l_0-norm, respectively. Therefore based on the heuristic approximations in Eq. (2.13) and Eq. (2.14), the highly non-convex objective function in Eq. (2.12) can be relaxed by replacing $rank(\cdot)$ with the nuclear norm (i.e., sum of the singular values: $\|\cdot\| = \sum_{i=1}^{M} \sigma_i(\cdot)$) and replacing the l_0-norm with l_1-norm (i.e., the sum of the absolute values of matrix entries: $\|\cdot\|_1 = \sum_{ij} |\cdot_{ij}|$), respectively.

And afterwards the relaxation yields a new convex optimization problem: minimization of the nuclear norm and l_1-norm, as shown in Eq. (2.15). This is the tightest convex relaxation of Eq. (2.12).

$$\min \|B\|_* + \lambda \|F\|_1 \quad s.t. \quad I = B + F \tag{2.15}$$

Solving this convex relaxation version is equivalent to solving the original low-rank matrix approximation problem if the condition that the rank of B to be recovered is not too large. The key point is how to solve the convex optimization, as expressed in of Eq. (2.15). This is a problem of Robust Principal Component Analysis (RPCA), several recovery algorithms have been proposed to solve this problem, such as the Augmented Lagrange Multiplier Method (ALM) [29], Accelerated Proximal Gradient [36], Singular Value Decomposition (SVD) [32], and so on.

2.3 Chapter summary

This chapter introduces the background knowledge. It first introduces the key points under the theory framework of compressive sensing which include signal sparsity, compressive measurements and signal reconstruction. And then the basic introduction to low-rank and sparse matrix decomposition is presented.

3 Multi-image fusion

3.1 Related work

W. Cao et al. [22] proposed Principal Component Analysis Transform (PCAT) and Wavelet Packet Transform (WPT) for remotely sensed image fusion. Sveinsson et al [23] proposed cluster based feature extraction and data fusion in the wavelet domain. Mallat et al. [24] proposed that if the wavelet coefficients undergo a modification like coefficient merging or quantization, then the inverse transform preserves this modification because the transform is non-redundant. Wen et al. [26] presented the relationships amongst image fusion methods and aimed to reveal the nature of various methods. Garzelli et al. [25] explained possibilities and limitations to use wavelets in image fusion. Leung et al. [27] proposed image fusion techniques using entropy. Milad et al. [28] presented a hybrid image fusion scheme that combines features of pixel and region based fusion, to be integrated in a surveillance system.

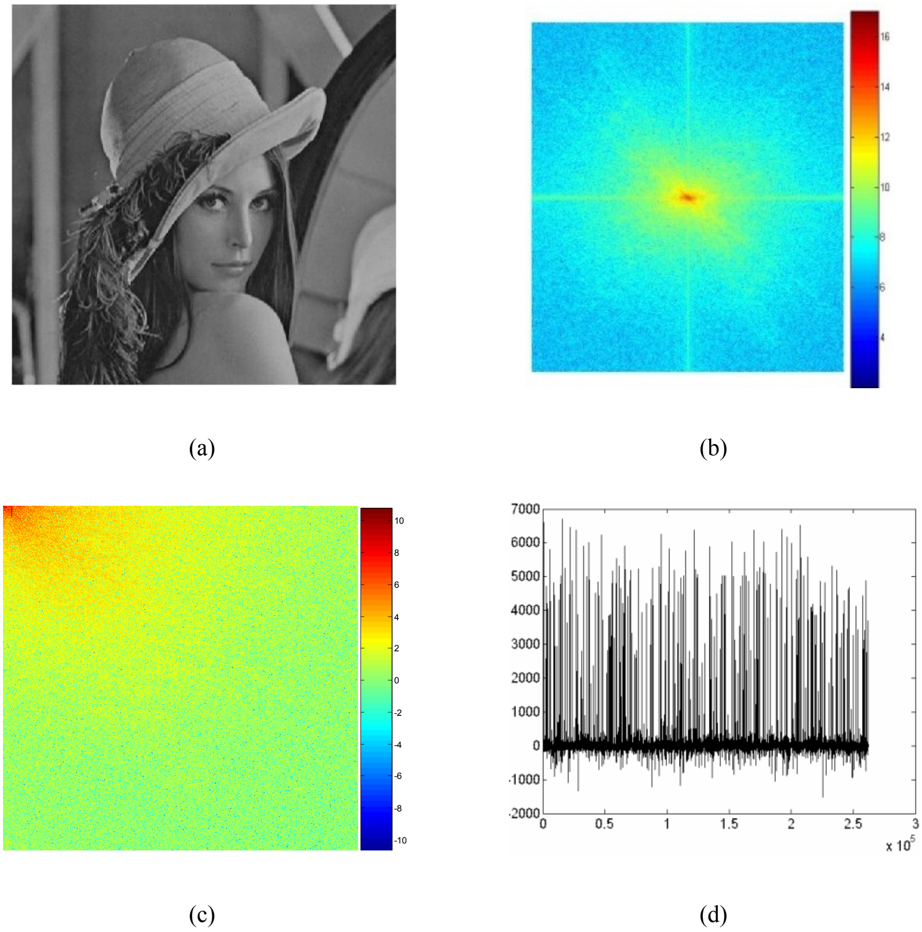
Regarding image fusion in the framework of CS, one natural way is to fuse the images after being reconstructed from the random projections. However, in order to reduce the computational complexity and to save storage space, a better way is to directly combine the measurements in the compressive domain, and then to reconstruct the fused image from the fused measurements. There are several different methods which have been proposed in recent years, such as a simple maximum selection fusion rule [80] and a weighted average based on entropy metrics of the original measurements [81].

In image compression, due to its computational simplicity and the fact that the spectral coefficients are real numbers, the Discrete Cosine Transformation (DCT) rather than the Fast Fourier Transformation (FFT) is widely used to represent a signal sparsely. The advantage of dealing with real rather than complex numbers also simplifies the algorithmic implementation of compressive approaches conceptually.

3.2 **Fusion scheme**

3.2.1 Image sparse representation

Sparse representations of images that have attracted considerable interest describe signals based on the sparsity and redundancy of their representations. For a natural image, the image data can be mapped to a sparse vector via a sparsifying transform. Different types of images have sparse representations under different transforms. Real-world images are known to have a sparse representation in the FFT, DCT and wavelet transform domain. The digital image "Lena" and its frequency transforms are shown in Fig.3.1.



(a)

(b)

(c)

(d)

Figure 3.1 (a) Original image. (b) Its FFT on log-scale. (c) Its DCT on log-scale. (d) Wavelet coefficients.

In Fig.3.1 (a), nearly all pixels values in the original image are non-zero. However, the image tends to concentrate its energy in the frequency domain, where most energy concentrates at low frequencies or at a few large coefficients. Fig.3.1 (b) shows the FFT of this original image on log-scale with shifting the zero-frequency component to the center of the image. The low-frequency components in an image are normally much larger in amplitude than the high-frequency components. The DCT relocates the compact energy in the upper left corner. Less energy or information is distributed over other areas, as shown in Fig.3.1 (c). The image is converted to a sparse vector in DCT domain. Most information of the original image is concentrated statistically in just a few large coefficients, while most of the high frequency coefficients are either zero or close to zero. Similarly, an image can be represented by just a few large coefficients in the wavelet transform domain, as shown in Fig. 3.1 (d). Thus, it can be said that the image has the sparsity property with a few large coefficients carrying most information using some orthogonal basis.

3.2.2 Sampling

For the reconstruction of the fused image, we first construct a viable sensing matrix which must satisfy the RIP. There are two ways to achieve this: (1) directly construct the sensing matrix to follow this property, and (2) reduce the problem to a known matrix that satisfies the RIP. Examples are the random Gaussian matrix, the uniform Spherical ensemble, Random Partial Fourier matrices, Toeplitz matrix, and so forth.

The random partial Fourier matrix is used to expand the applicability of compressive sensing to large scale data such as 2D images due to the special structure of the Fourier transform under the partial Fourier ensemble [82]. Inspired by this work, the thesis proposes a sampling model to account for the property of the DCT in the frequency domain as shown in Fig.3.2. The DCT relocates the energy of a digital image in the frequency domain. Most of the energy of a digital image concentrates at low frequencies (upper left corner shown in Fig. 3.1(c)). Hence most information of an image can found in the measurements located at the upper left corner of the image in the DCT domain. The sampling model contains many radial lines extending from the upper left corner to the other side of an image, as shown in Fig.3.2. The measurement matrix then is constructed from the sampling pattern in the 2D discrete cosine plane created by using nearest neighbor techniques.

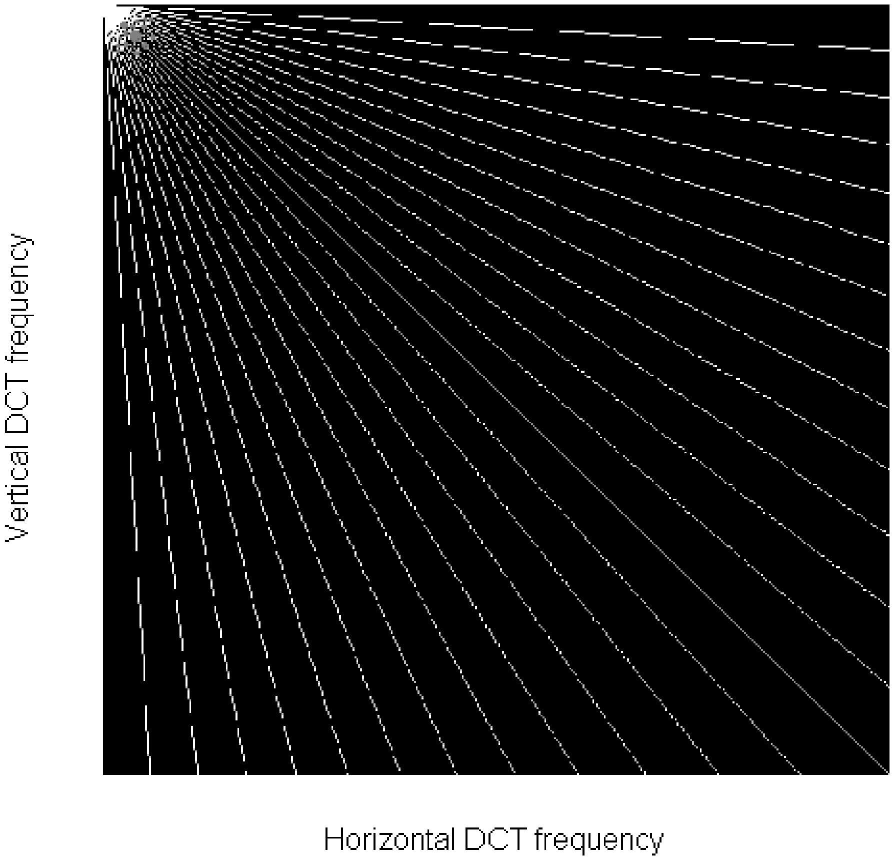


Figure 3.2 DCT based sampling model (zero frequency in the upper left corner).

3.2.3 Fusion

For most of the conventional fusion approaches, the image fusion is performed on the level of the source images. With the emergence of CS theory, however, the fusion progress can be implemented in the compressive domain. That is, we first combine the individual linear measurement of multi-input images into a single composite measurement, and then reconstruct the fused image from the composite measurement.

Consider a natural image with the size of $n \times n$ pixels, we usually stack the image data into a one dimensional column vector f of length $N(N = n \times n)$ for the purpose of simplifying the complexity of the computation, It has already been known that the column vector f is sparse under the orthogonal 2D DCT basis Ψ according to Fig.

3.1(c), that is, $f = \Psi x$, where x is the K-sparse coefficients. The measurement vector y is then considered as the projection of images onto the column vectors of the measurement matrix Φ. Mathematically speaking, the relationship can therefore be expressed as $y = \Phi f = \Phi\Psi x = \hat{\Phi} x$.

There are P images with the same size $n \times n$ pixels are supposed to be fused and all the images have similar spectral features. Each image matrix is transformed to one dimensional column $f_p (p = 1, 2, ..., P)$. All the vectors have compact representations in terms of the significant coefficients in the orthogonal basis. We collect the corresponding linear measurements $y_p \left(p = 1, 2, ..., P\right)$ with length $M\left(M < N\right)$ in one large augmented observation vector of size $M \times P$ (rather than $N \times P$). Hence, the measurements are not the simple pixel values of the original images any more. The fusion among the original images can be considered naturally as the fusion among the linear measurements of $y_p \left(p = 1, 2, ..., P\right)$, which contain the important information reflecting the image texture.

Multi-Scale wavelet decomposition shows remarkable advantages in the representation of a signal. In the fusion scheme, we apply a single-level one dimensional Daubechies wavelet transform to decompose the linear measurement vectors into two components: the approximation coefficients $A_p \left(p = 1, 2, ..., P\right)$ and the detail coefficients $D_p \left(p = 1, 2, ..., P\right)$. As the larger a coefficient is, the more information it carries, a weighted mean is applied to incorporate the contributions of all inputs so that data elements with a high weight contribute more to the weighted mean than elements with a low weight. The fused approximation coefficient A and detail coefficient D can be formulated as:

$$A = \sum_{p=1}^{P} \alpha_p A_p \quad , \tag{3.1}$$

$$D = \sum_{p=1}^{P} \beta_p D_p \quad , \tag{3.2}$$

where α_p and β_p are the weighting factors corresponding to approximations and details, respectively, and defined as

$$\alpha_p = \frac{|A_p|}{\sum_{p=1}^{P} |A_p|}, \quad \beta_p = \frac{|D_p|}{\sum_{p=1}^{P} |D_p|} . \tag{3.3}$$

Consequently the fused linear measurement y is obtained through the inverse discrete wavelet transform. This is a process by which components can be assembled back into the original signal without loss of information. Finally, the one dimensional column vector f of the fused image is reconstructed from the fused linear measurement y via the recovery algorithm total variation minimization.

3.3 Experimental results

In this section, we perform two groups of comparisons for the performance evaluation to illustrate the effectiveness of the proposed approach. In the experiments all the input images have the sparsity property in the 2D discrete cosine transform domain. The fused images are reconstructed from measurements. In the work, we compare the proposed scheme with the maximum selection fusion rule proposed in [80] and the block-based weighted average fusion rule presented in [81].

In the first group, the comparison is performed on a pair of multi-focus images with size of 512×512 pixels. We take the classical "Lena" image as a reference image, as shown in Fig. 3.3 (a). We artificially produce a pair of out-of-focus images, as shown respectively in Fig. 3.3 (b) and Fig. 3.3 (c). Blurring is accomplished by using a Gaussian low-pass filter. The fusion results using the maximum selection fusion, weighted average fusion and our method are shown in Fig.3.3 (d), (e) and (f), respectively.

In the second group, multi-modal medical images are used as input. The first one is a Computed Tomography (CT) image shown in Fig.3.4 (a) and the other one is a Magnetic Resonance Image (MRI), see Fig.3.4 (b). The fusion results using the maximum selection fusion, weighted average fusion and our method are shown in Fig.3.4 (d), (e) and (f), respectively.

It is well known that assessing image fusion performance in a real application is a complicated issue. In many cases qualitative criteria such as visual analysis is used to assess the fusion result. However, a more accurate and reliable evaluation is to combine visual assessment based on a subjective qualitative analysis with a parameter assessment based on an objective quantitative analysis. To evaluate our proposed algorithm perceptually is first conducted, and afterwards we use several quality measures to compare its results to previous approaches.

3.3.1 Perceptual quality evaluation

Perceptual evaluation mainly assesses the visual quality of the fused image by means of observing and contrasting details. Based on a visual comparison, the fusion results of the proposed method shown in Fig.3.3 (f) contain most of the details of the individual input images shown in Fig.3.3 (b) and Fig.3.3 (c). On one hand the image shown in Fig.3.3 (f) looks smoother than image in Fig.3.3 (d), on the other hand it is clearer than Fig.3.3 (e).

With regard to the visual comparison of the second group, the fusion result of proposed method shown in Fig.3.4 (e) contains more information than the input images in Fig.3.4 (a) and (b). Fig.3.4 (e) has more details than Fig.3.4 (c), whereas Fig.3.4 (e) has a higher contrast than the image in Fig.3.4 (d). For a comparison of the image details the enlarged fusion results for all methods are shown in Fig. 3.5.

This approach outperforms the method of maximum selection fusion and weighted average fusion when judging the perceptual quality of the fusion results for both image sets.

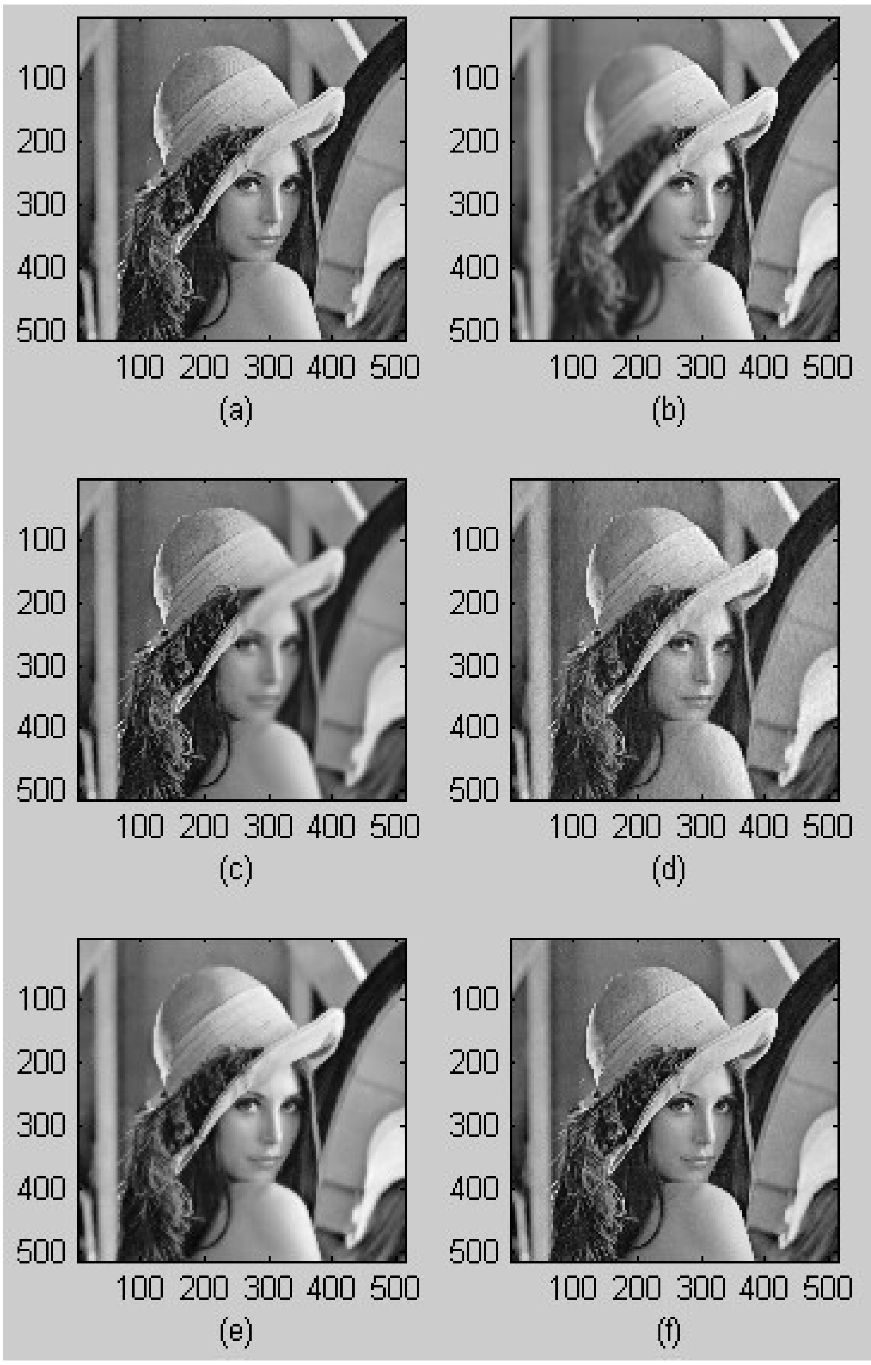


Figure 3.3 (a) Reference image. (b) Focus on the left part. (c) Focus on the right part. (d) Fusion result using maximum selection. (e) Fusion result using weighted average (f) Fusion result of proposed method.

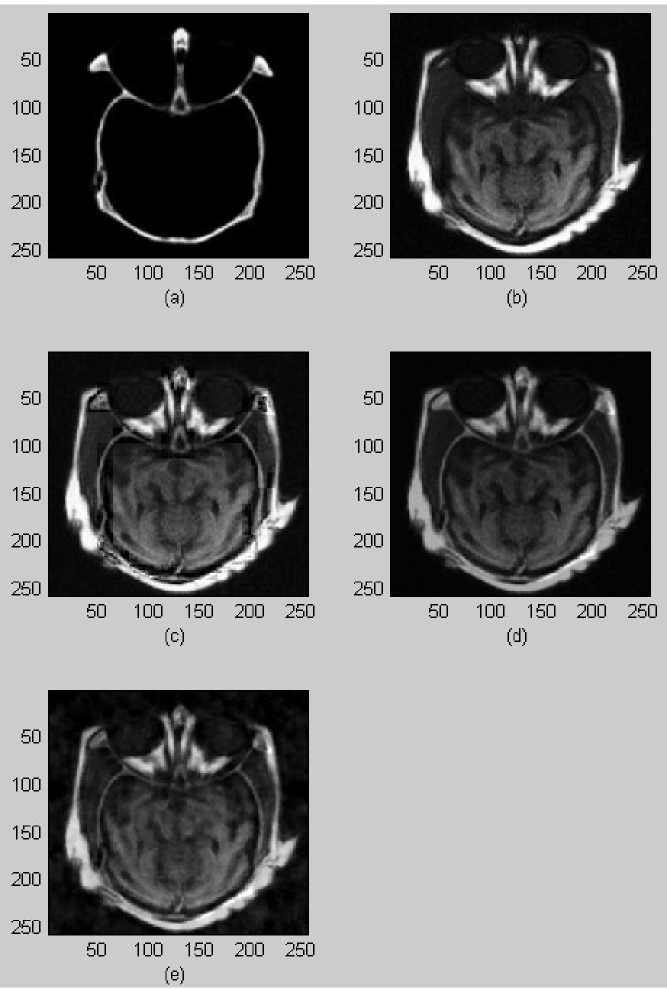


Figure 3.4 (a) CT image. (b) MRI image. (c) Fusion result using maximum selection. (d) Fusion result using weighted average. (e) Fusion result of proposed method.

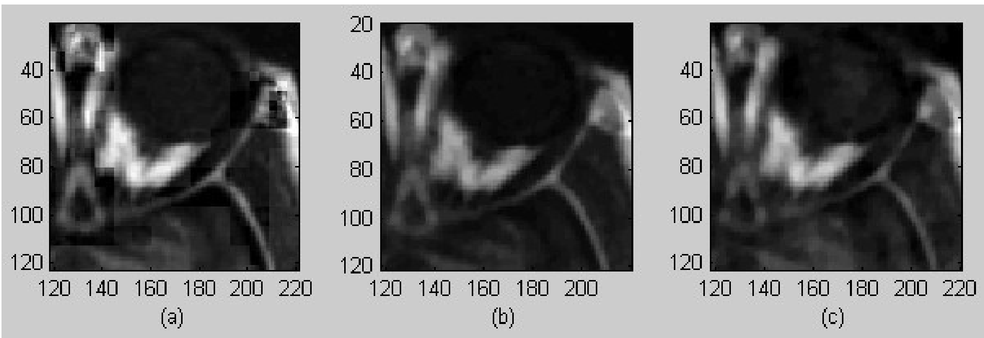


Figure 3.5 Images in zoom in view. (a) Fusion result using maximum selection. (b) Fusion result using weighted average. (c) Fusion result of proposed method.

3.3.2 Objective quantity evaluation

In general, there are a few quality measures that are commonly used to evaluate image fusion results: image entropy, mutual information and average gradient.

(a) Image entropy (IE)

Image entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. For an 8-bit single channel image, the image entropy is defined as:

$$H = -\sum_{i=0}^{255} P_i \log_2 P_i \quad . \tag{3.4}$$

where P_i is the probability of gray level i in the evaluated region and it is approximately given by

$$P_i = \frac{f_i}{N} \ , \tag{3.5}$$

where f_i is the frequency of gray level i and N denotes the total number of pixels in the image. The higher the value of the image entropy is, the more textural information is contained in the fused image.

(b) Mutual information (MI)

Mutual Information is often used to evaluate image fusion quality. Let the joint histogram of source image $A(B)$ and the fused image F be $p_{FA}(f,a)\left(p_{FB}(f,b)\right)$. Then the mutual information between the source image and the fused image is given by

$$I_{FA}(f,a) = \sum_{f,a} p_{FA}(f,a) \log_2 \frac{p_{FA}(f,a)}{p_F(f)\, p_A(a)} \ , \tag{3.6}$$

$$I_{FB}(f,b)=\sum_{f,b}p_{FB}(f,b)\log_2\frac{p_{FB}(f,b)}{p_F(f)p_B(b)}. \tag{3.7}$$

The image fusion performance can be measured by:

$$MI_F^{AB}=I_{FA}(f,a)+I_{FB}(f,b), \tag{3.8}$$

where larger values imply better image quality.

(c) Average gradient (AG)

The average gradient is a measure of contrast in a photographic image. It is sensitive to reflect the image of the tiny details contrast. It is commonly used to evaluate the clarity of image. We use average gradient as a criterion for image fusion quality. The greater the average gradient value is, the sharper is the image. It can be calculated as:

$$\bar{g}=\frac{1}{n}\sum\sqrt{\frac{(\Delta I_x)^2+(\Delta I_y)^2}{2}}, \tag{3.9}$$

where n is the size of the image, ΔI_x and ΔI_y are the differences in horizontal and vertical direction, respectively.

The performance assessments of the fusion results shown in Fig.3.3 and Fig.3.4 based on the defined criterions (i.e., IE, MI and AG) are listed in Tables 3.1 and Table 3.2.

Experiment 1

Regarding the "Lena" image shown in Figure 3.3 (a), for the purpose of comparing mutual information parameter in detail, we calculate not only the mutual information between the fused image and the individual image, but also the mutual information between the fused image and the original reference image as listed in Table 3.1. I_{FA} denotes the mutual information between the fused image and the source image A, while I_{FB} is the mutual information between the fused image and the source image B. MI is the sum of I_{FA} and I_{FB}. I_{FR} is the mutual information between the fused image and reference image. We present here the value using two decimal places in the Table 3.1.

Table 3.1 Quantitive evaluation of the multi-focus images shown in Fig.3.3.

Methods	Performance Evaluation Measures					
	IE	I_{FA}	I_{FB}	MI	I_{FR}	AG
Proposed method	7.12	2.75	2.85	5.60	3.03	2.86
Maximum selection	7.10	2.40	2.57	4.97	2.81	3.59
Weighted average	6.99	2.68	2.71	5.39	3.02	2.21

It can be seen from Table 3.1 that the proposed method outperforms the other methods in terms of IE and MI, which means that the fusion result of proposed method contains more details than those of the other methods. The visual comparison above also suggests that the fusion result of proposed method is superior to the result of the maximum selection method and clearer than the result of the weighted average method, though the average gradient value for the maximum selection method is a little bit larger than that for proposed method. Overall, based on the visual comparison and comparison using objective measures, we can draw the conclusion that proposed method achieves better performance than the other two methods.

Experiment 2

Regarding the medical image, we only compare the three performance assessment measures (IE, MI and AG), since we do not have the reference image. The results are shown in Table 3.2. It can be seen easily that proposed method performs better than the other two methods when comparing the IE and MI results in Table 3.2. Taking the visual analysis into account, we conclude that proposed method outperforms the methods of maximum selection fusion rule and average gradient fusion rule.

Thus, by considering the qualitative analysis and the quantitative evaluation it is concluded that the results of the proposed fusion scheme are superior when compared to the maximum selection fusion rule and the weighted average fusion rule.

Table 3.2 Quantity evaluation of multi-modal images in Fig.3.4.

Methods	Performance Evaluation Measures		
	IE	MI	AG
Our Method	6.9763	5.2867	5.1054
Maximum selection	6.6992	5.1544	6.5336
Weighted average	5.8196	3.7439	3.0164

3.4 Chapter summary

Compressive sensing provides a novel framework to acquire and reconstruct a signal or digital image from sparse measurements acquired at sub-Nyquist sampling rate. In this chapter, an effective image fusion scheme based on a DCT sampling model for compressive sensing imaging is presented first. A sparse sampling model according to the DCT-based spectral energy distribution is proposed. The compressive measurements of multiple input images obtained with the proposed sampling model are fused to a composite measurement by combining their wavelet approximation coefficients and their detail coefficients separately. The combination is done by applying a weighting operation for every sampling location according to the statistical distribution. Furthermore, the fused image is reconstructed from the composite measurement by solving a problem of total variation minimization. The computational complexity decreases due to the fact that the proposed scheme only needs incomplete measurements rather than acquiring all the samples of the whole image. Moreover, although our method performs the fusion in the sparse domain, it preserves much richer texture information of the individual input images compared with other fusion schemes. Experiments demonstrate the promising performance of the proposed approach.

4 Super-resolution of depth image

4.1 Range/Depth imaging

The conventional image sensors such as CCD/CMOS measure the intensity or color image (2D image) information of the scene. 2D images are of limited use in terms of estimation of surfaces due to lack the depth information of the scene and therefore pixel values are indirectly related to surface geometry.

In the past years the range imaging technology became more and more attractive to a growing research community due to that it is powerful to provide range information (depth map). The technology collects mainly range values between the imaging sensor and the points of the object in a scene.

Range images that are also referred to as depth images, depth maps, xyz maps, surface profiles. Range images are a special class of digital images in which each pixel expresses the distance between a known reference frame and a visible point on object surface in the scene. In general, range images can be represented in two basic forms. The one is a matrix of depth values of points along the directions of the x, y image axes, which makes spatial organization explicit. The other one is a list of three dimensional (3D) coordinates in a given reference frame (cloud of points), for which no specific order is required. Thus, a range image can reproduce the 3D structure of a scene.

Range images are acquired with range sensors. A range imaging sensor is any combination of hardware and software capable of producing a range image of a real-world scene under appropriate operating conditions. It collects large amounts of 3D coordinate data from visible surfaces in a scene and can be used in a wide variety of applications. It is a unique imaging device that is sometimes referred to as a *range camera* in which the image data points explicitly represent scene surface geometry as samples points. The optical range imaging sensors normally used in computer vision can be classified into two categories: active and passive. The active ones transmit some form of energy into a scene to receive a return signal that allows determining ranges. While instead of project any form of energy into a scene, the passive ones use naturally present light to obtain range data in single shots or multi-frame grabs and work similarly to single-lens-reflex or film cameras. Many different technologies can be used to build 3D scanning devices, to some extent each technology more or less comes with its own limitations, advantages and costs. Some of different techniques are presented as the following:

Structured light

It is a non-contact active triangulation technology for 3D range measurement. Structure light 3D scanning is about determining the 3D structure of a scene based on the distortion of the projected pattern. Structured light 3D scanner is a device for measuring the 3D shape of an object using projected light patterns and a camera system. It projects a pattern of light on the subject and looks at the deformation of the pattern on the subject. The pattern is projected onto the subject using either an LCD projector or other stable light source. A camera, offset slightly from the pattern projector, looks at the shape of the pattern and calculates the distance of every point in the field of view [112].

As the principle shown in Figure 4.1 [75], projecting a narrow band of light onto a three-dimensionally shaped surface produces a line of illumination that appears distorted from other perspectives than that of the projector, and can be used for an exact geometric reconstruction of the surface shape (light section). A faster and more versatile method is the projection of patterns consisting of many stripes at once, or of arbitrary fringes, as this allows for the acquisition of a multitude of samples simultaneously. Seen from different viewpoints, the pattern appears geometrically distorted due to the surface shape of the object. Although many other variants of structured light projection are possible, patterns of parallel stripes are widely used. The figure shows the geometrical deformation of a single stripe projected onto a simple 3D surface. The displacement of the stripes allows for an exact retrieval of the 3D coordinates of any details on the object's surface [75] [76].

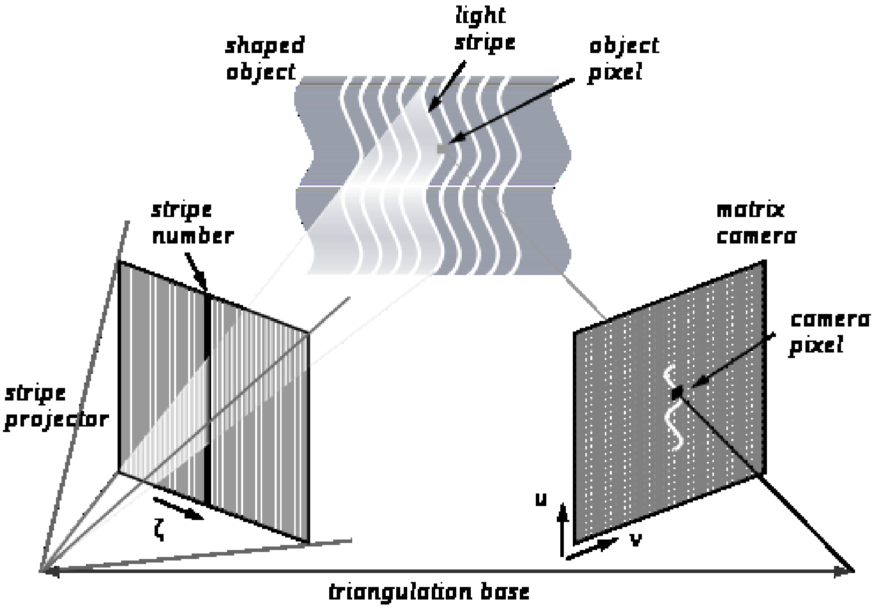


Figure 4.1 Triangulation principle shown by one of multiple stripes

The main advantage of structured-light 3D scanners is speed and precision. Instead of scanning one point at a time, structured light scanners scan multiple points or the entire field of view at once. Scanning an entire field of view in a fraction of a second generates profiles that are exponentially more precise than laser triangulation. This reduces or eliminates the problem of distortion from motion. Some existing systems are capable of scanning moving objects in real-time. However, depending on the specific applications, the main drawbacks of this technology include missing range data at region of the scene which are not visible to the light projector and are visible to 2D camera or vice versa [61]. A real-time scanner using digital fringe projection and phase-shifting technique (a various structured light method) was developed to capture, reconstruct, and render high-density details of dynamically deformable objects (such as facial expressions) at 40 frames per second [62].

Stereoscopy

Stereoscopy, sometimes called stereoscopic imaging, is a passive technique for creating or enhancing the illusion of depth in an image by means of stereopsis for binocular vision [113]. The basic technique of stereoscopy is to present offset images that are displayed separately to the left and the right eye. Both of these 2D offset images are then combined in the brain to give the perception of 3D depth. A typical approach of stereoscopy is computer stereo vision in which two cameras are displaced horizontally from each other, while another are used to obtain two differing views on a scene, in a manner similar to human binocular vision. By comparing information about a scene from two vantage points, the relative depth information can be extracted in the form of disparities which are inversely proportional to the differences in distance to the objects.

There are two main problems to conduct in stereo vision:

- The correspondence problem

  Given two or more images of the same 3D scene, taken from different points of view, the correspondence problem is to find a set of points in one image which can be identified as the same points in another image. To do this, try to match points or features from one image with the same points or features in another image. In this problem, the disparity map can be computed when the corresponding points are known.

- The reconstruction problem

  Then a 3D map of the scene is then reconstructed from the disparity map.

The stereo vision system has the advantages of being safe for human, they are cheap, there is in principle no limit to the distance that can be measured, and there is no interference. However, the main drawbacks include it has no range data in a uniform region, unavoidable triangulation errors, objects not appearing in image data of vision cannot be measured and difficulty to solve the occlusion problem [65].

<u>Laser Pulse Rangefinder</u>

The range finding using pulsed lasers is an active approach based on Time-of-Flight principle for measuring the distance of objects in the scene. As shown in Figure 4.2 [67], it typically consists of a laser pulse transmitter, the necessary optics, two receiver channels and a Time-to-Digital Converter (TDC). The pulse laser emits a short light pulse which starts the time measurement in the receiver. As soon as the pulse reflected from the object reaches the photo detector the time measurement is stopped. The elapsed time between start and stop pulse is used in TDC to compute the distance to the reflector [63] [64].

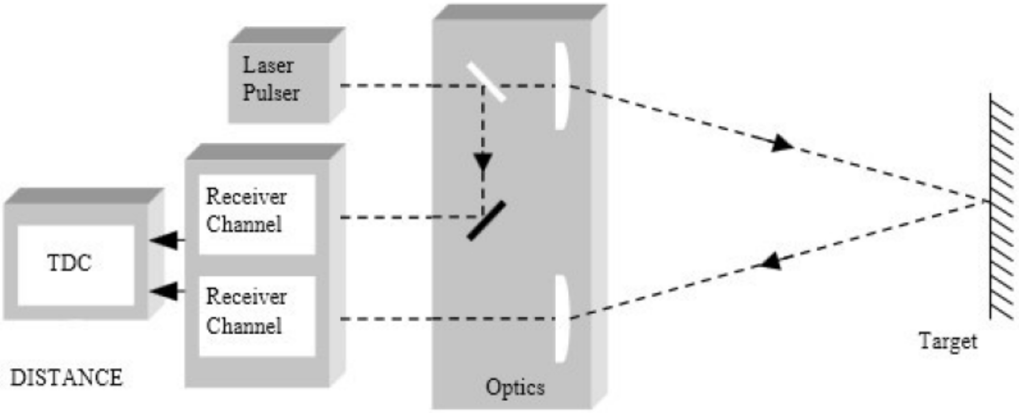


Figure 4.2 Principle of a pulsed time-of-flight laser rangefinder

Due to the high speed of light, this technique is not appropriate for high precision sub-millimeter measurements, where triangulation and other techniques are often used. The main drawback of laser range finders is their long acquisition time which is due to the scanning process. The output of laser range finders are point clouds which are not directly usable in most of 3D applications and therefore they should be converted to 3D models or range images which is itself a time consuming process.

Time of Flight

Recently the sensors had been developed that acquire distance information based ToF principle and for which distance errors based on different causes can be observed. There are several technologies present for range imaging such as the PMD working on modulated, incoherent infrared light or by using the depth imprint of an emitted light pulse by fast shuttering.

Range imaging in a 3D ToF camera is the fusion of the distance measurement technique with the imaging aspect. It consists of an optical transmitter and an optical receiver. The principle of the range measurement in a ToF camera is based on the measurement of the time that the light needs to travel from a target to a reference point, a detector. This so-called Time of Flight is directly proportional to the distance the light travels. The basic ToF principle is shown in Figure 4.3 [114].

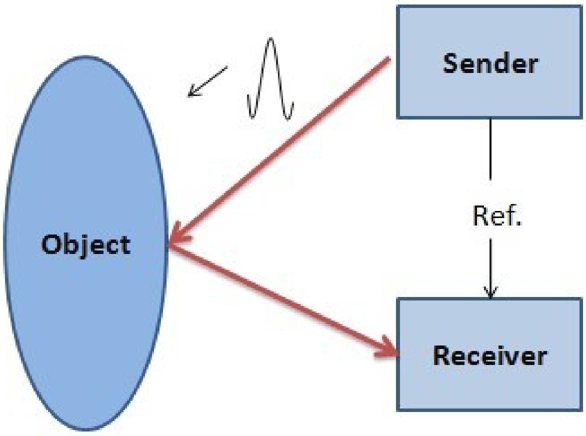


Figure 4.3 Basic Time-of-Flight principle.

In its most simple form, a light pulse is transmitted by a sender unit and the target distance is measured by determining the turn-around time that the pulse travels from the sender to the target and back to the receiver. If we use T to denote the echo time and c be the speed of light, with knowledge of the speed of light, the distance can then easily be calculated as Eq. 4.1.

$$D = \frac{T \cdot c}{2} . \tag{4.1}$$

In the work, we use a 3D ToF camera based on the Photonic Mixer Device (PMD) exploiting phase shift measurement. The entire scene is illuminated with modulated light. PMD technology allows us to observe this illuminated scene with an intelligent pixel array, where each pixel can individually measure the turnaround time of the modulated light. Typically this can be done by using continuous modulation and measuring the phase delay in each pixel [34].

We use a modulated light signal f as a light source. With four samples A_1, A_2, A_3 and A_4, each shifted by $\pi / 2$, the strength of the received signal (also termed as modulation amplitude) s and the gray scale value g are respectively formulated as Eq. (4.2) and Eq. (4.3) [35].

$$s = \frac{\sqrt{(A_1 - A_3)^2 + (A_2 - A_4)^2}}{2}, \tag{4.2}$$

$$g = \frac{A_1 + A_2 + A_3 + A_4}{2} \tag{4.3}$$
.

To calculate the phase delay $\Delta\varphi$, the autocorrelation function of the electrical an optical signal is analyzed by a phase-shift algorithm, we calculate the phase delay by using

$$\Delta\varphi = \arg\tan\left(\frac{A_1 - A_3}{A_2 - A_4}\right) \tag{4.4}$$
.

Knowing the phase shift to each pixel on the sensor, the sensor directly measures the distance D_{pixel} to the captured object [35]:

$$D_{pixel} = \frac{\Delta\varphi \cdot c}{4\pi \cdot f_{\text{mod}}} \tag{4.5}$$
.

Advantages and Limitations of 3D Range Imaging

Many researchers have found advantages and limitations of range imaging prototype sensors and especially the influence of an ambient environment.

Some of the advantages of 3D range image cameras are:

- Delivering range, amplitude and intensity maps in one frame at the same time
- Capturing static and dynamic scenes
- Ease of use at day and night
- Insensitivity to background light changes in indoor environments

- Reasonable price

While currently the main limitations to 3D range imaging cameras are:

- Low lateral resolution

- Ambient environment influencing measurements

- Noisy range data from poorly reflecting surfaces

As 3D range imaging cameras have their own strengths and limitations, lots of approaches have been proposed to combine sensors of difference modality in order to take advantages of more sensors and perform in an optimal way.

4.2 Related work

3D depth sensing is a key component of many machine vision systems. Among existing technologies, a new kind of camera developed in recent decade measures the time-of-flight of infrared light between camera and object in a scene to reconstruct 3D scene geometry at real-time frame rates, largely independently of scene texture. However, being a relatively young technology, the state-of-the-art ToF sensors have not enjoyed the same advances with respect to spatial image resolution, image quality, and acquisition speed, which have been made in traditional 2D CCD image sensors. In consequence, current ToF sensors provide depth readings of comparably low image resolution (e.g. typically 64×48 up to 204×204 pixels for the PMD based 3D ToF cameras). And therefore the ability to capture not only good quality but also high-resolution depth image is desired. Attempts have been made to enhance the resolution of depth map.

The depth accuracy of ToF sensor can be increased by a variety of methods, e.g. by accounting for ambient light [98], simulating the shape of the reflected signal [99], and applying time gated super-resolution [100]. While these methods improve resolution in the depth direction, they are not directly related to improving resolution in X-Y plane. Increasing the number of pixels per unit area (i.e., reduce the pixel size) by sensor manufacturing techniques is an option to capture High-Resolution (HR) depth image. However, to some extent, it is difficult due to physical constraint. Another approach for enhancing the X-Y resolution of image is to increase the chip size, which leads to an increase in capacitance. The high cost for high precision optics and imaging sensors is also a significant concern in a lot of commercial applications regarding HR imaging.

Another way to address the problem is to use signal processing to post process the captured images, to trade off computational cost with the hardware cost. These techniques are specifically referred as Super-Resolution (SR) reconstruction. The major advantage of the signal processing approach is that the existing low-resolution imaging systems can be still used and therefore apparently it may cost less in comparison with *physical* approach mentioned before. SR of depth image can be broadly categorized into two classes: (1) approach with help of high-resolution color image; (2) approach based on only depth map.

(1) Approaches with help of high-resolution 2D image

Super-resolution of depth image has been accomplished by the operation of fusion or combination with a high resolution color image acquired in the same scene from the same location. A common up-sampling recipe enforces heuristics between depth and intensity images, It relies on the co-occurrence of depth and intensity discontinuities, on depth smoothness in areas of low texture, and careful registration for the object of interest. The low resolution depth map can be up-sampled and regularized subject to an edge consistency term concerning color image. An approach that put this idea into practice was proposed by Diebel et al. [43]. They fused depth maps of a laser range finder with intensity maps of a color camera by defining the posterior probability of the high-resolution reconstruction as a Markov Random Fields (MRF) and optimizing for the Maximum-A-Posterior (MAP) solution. Following the similar way, Kopf et al. [44] proposed an alternative fusion method called joint bilateral up-sampling. Their algorithm utilized a modification of the bilateral filter, an edge-preserving smoothing filter for intensity images [46]. The bilateral filter locally shapes the spatial smoothing kernel by multiplying with a color similarity term, as known as the range term, which yields an edge-preserving smoothing filter. Kopf et al. capitalized on this adaptive smoothing capability and bilaterally up-sampled a low-resolution tone mapping result such that it matches the resolution of a multi-megapixel intensity image. Recently, Crabb et al. [50] applied bilateral upsampling to range data captured by a time-of-flight camera in order to do real-time matting. Yang et al. [45] also proposed an upsampling technique based upon a bilateral filter. However, they rather used the high resolution image to create a cost volume to which they apply a standard bilateral filter than used a joint bilateral technique to link the two images. This required them to run a 2D bilateral kernel over multiple slices of a volume. Park et al. [49] improved on these results with better image alignment, outlier detection and also by allowing for user interaction to refine the depth. Incorrect depth estimates can come about if texture from the intensity image propagates into regions of smooth depth. The joint bilateral filter performs edge-preserving smoothing to a low resolution depth map with assuming the occurrences of edges are highly correlated between the dept map and the color image. Although these approaches can reproduce high frequency detail, they incorrectly supposed that color is correlated with depth. This could lead to difficulties with colored textures and when a true depth discontinuity is invisible in the color channel.

Markov Random Field

The first successful attempt to up-sample depth values to match the resolution of a color image was based on Markov Random Field (MRF) that used color information from a color image, and depth map where available [43]. The terms of the MRF energy function attempt to enforce depth smoothness, but allow for depth discontinuities across color borders. The belief underlying the method is that areas of constant color are most likely areas of constant depth. Hence, the depth at any given pixel is mostly similar to that of its neighbors that are within the same color boundary.

A posterior probability of each pixel value in a high-resolution depth map is defined as $P(H \mid L)$, where L is the vector of the observed pixel values L_{pl}'s in a low resolution depth map and H denotes the vector of the random variables for the pixels values H_{pl}'s in a high resolution depth map. According to Bayes' rule, $P(H \mid L)$ can be denoted as the product of the likelihood probability $P(L \mid H)$ and a priori probability $P(H)$. Therefore an optimal HR depth map can be obtained by finding the configuration of H which maximizes $P(H \mid L)$.

Maximum a posterior probability can be obtained by minimizing the following energy function

$$E(H) = D(H) + \mu V(H) \tag{4.6}$$

.

The data penalty function $D(H)$ makes H_{pl} and L_{pl} similar, and can be expressed as

$$D(H) = \sum_{pl} \left(L_{pl} - H_{pl} \right)^2 \tag{4.7}$$

.

The smoothness function $V(H)$ is defined as

$$V(H) = \sum_{p} \sum_{q \in N(p)} e^{-\alpha \left\| C_p - C_q \right\|^2} \left(H_p - H_q \right)^2 \tag{4.8}$$

,

where $N(p)$ denotes the neighborhood of p and α is constants. The smoothness function basically enforces $H_p$and $H_q$to be similar each other. However, in the cases of large differences of $\left\| C_p - C_q \right\|^2$ in a color image, the smoothness term becomes small and therefore preserves the edges in the depth image by admitting different values of H_p and H_q.

Joint Bilateral Upsampling (JBU)

The bilateral filter is an edge preserving smoothing filter which adaptively changes a spatial kernel for smoothing based on the intensity differences of an input image [46]. The output value of the filtering at a pixel is computed as the weighted average intensity for the neighboring pixels. However, the intensity difference between a pixel and its neighboring one controls the weight for averaging and therefore the edges can be preserved while the non-edge regions are smoothed.

As it is known, the Joint Bilateral Filter (JBF) is a modified version of the bilateral filter. It employs a LR depth map and a HR color image together to up-sample the depth map. The up-sampling techniques based on JBF assume the occurrences of edges between depth and color images are highly correlated [44]. The filtered value H_q at the pixel p in a HR depth map can be denoted as

$$H_p = \frac{1}{k_p} \sum_{q' \in \Omega} L_{q'} f\left(\left\|p' - q'\right\|\right) g\left(\left\|C_p - C_q\right\|\right) \quad (4.9)$$
,

where k_p is a normalizing term. Ω is the neighborhood of p'. L_q denotes the pixel value at q' in a LR depth map. p' and q' are the pixels in a LR image which correspond to the pixels p and q in a HR image, respectively. C_p and C_q are the pixel values at p and q in a HR color image. $f(\cdot)$ and $g(\cdot)$ are the weights of the spatial term and the range term, respectively, which are 2D Gaussian kernels with different means and variances. The spatial term is assigned a large weight when $\left\|p' - q'\right\|$ becomes smaller. Thus, H_p is basically the average intensity for the neighboring pixels resulting in a smoothing effect. However, the range term performs smoothing adaptively in the up-sampled depth image using the information of color image.

(2) Approaches based on only Depth Map

Interpolation-based methods for single depth image

Interpolation has been widely used in many image super-resolution applications due to the fact that it is simple and easy to implement. Interpolation-based methods generate a HR image from its LR version by estimating the pixel intensities on an up-sampled grid [47][48]. However, this method tends to blur the high frequency details. Most generally, the simplest up-sampling techniques use nearest-neighbor, bilinear, or bicubic interpolation to determine image values at interpolated coordinates of the input domain. Such increases in resolution occur without regard for the input's frequency content. As a result, nearest-neighbor interpolation turns curved surfaces into jagged steps, while bilinear and bicubic interpolation smooth out sharp boundaries. Such artifacts are hard to measure numerically, but can be perceptually quite obvious both in intensity and depth images.

Methods based on multiple depth images:

The SR problem traditionally focuses on fusing multiple LR observations to reconstruct a higher resolution image [51]. Schuon et al. [52] applied combination of multiple LR depth images with different camera centers in a framework of optimization which is designed to be robust to the random noise characteristics of ToF sensors. In order to reduce the noise in each individual depth image, Yang et al. [45] composited together multiple depths from the same viewpoint to make a “single” depth image for further super-resolution, and Rajagopalan et al. [96] used an MRF formulation to fuse together several low resolution depth images to create a final higher resolution image. Fusing multiple sets of noisy scans has also been demonstrated for effective scanning of individual 3D shapes [56]. Hahne et al. [54] combined combine depth scans in a manner similar to exposure-bracketing for high dynamic range photography. Using Graphics Processing Unit (GPU) acceleration, Izadi et al. [55] made a system which registers and merges multiple depth images of a scene in real time.

4.3 **Proposed strategy**

4.3.1 Analysis of problem model

The SR algorithms attempt to generate a single HR image from one or more LR images of the same scene. The goal is to reconstruct the high-frequency missing information in one way that approximates the desired HR image as closely as possible.

There are both single-image and multiple-images variants of SR multiple-images based SR algorithms utilize the sub-pixel shifts between multiple low-resolution images of the same scene [57]. They create an improved resolution image by fusing information from all LR images. However, how to recover missing information from a single LR image is more interesting and challenging. That is to say, the problem of single-image SR is particularly important because in our application only a single, LR depth map is available and the up-sampling technique must be applied as a post-processing. It is our goal to obtain high-resolution depth map of a static scene despite the significant noise in the raw data. We enhance X-Y resolution of depth image and meanwhile reduce the overall random noise level by performing SR technique.

The first step to comprehensively analyze the SR image reconstruction problem is to formulate an observation model that relates the original HR image to the observed LR images. In the process of recording a depth image, there is a natural loss of resolution caused by the optical distortions, motion blur, and insufficient sensor density [101]. Thus, the recorded image denoted by f usually suffers from effects of warping denoted by I , blurring denoted by G and down-sampling operators denoted by D. In addition, assuming that LR image is corrupted by additive noise g , then the problem model can be observed as:

$$y_k = I_k D_k G_k f + \varepsilon_k \tag{4.10}$$

where k denotes the k^{th} LR image observed from the recorded HR image. The common observation model is shown in Figure 4.4.

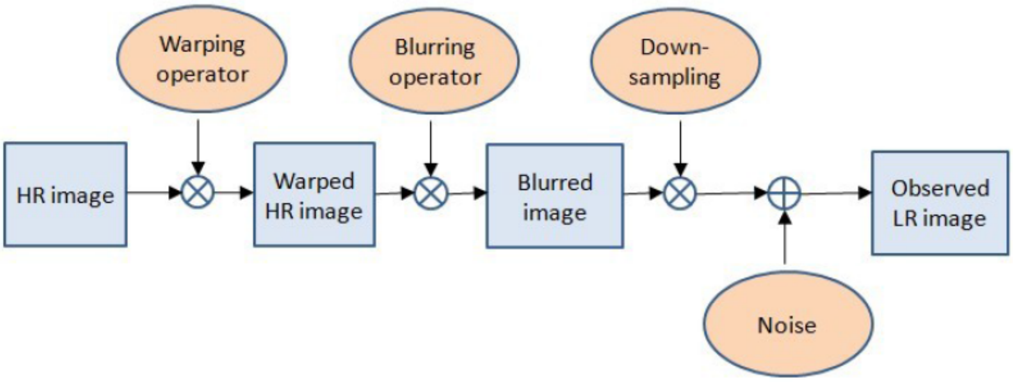


Figure 4.4 Problem model relating LR image to HR image.

Considering that the depth images are obtained after the processes of distortion correction and noise removal and in consequence I is an identity matrix, the problem model can be represented as:

$$y_k = D_k G_k f \, . \tag{4.11}$$

Let the size of desired HR image $C_1N_1 \times C_2N_2$ is denoted as the vector $f = [f_1, f_2, \ldots, f_N]^T \left(N = C_1N_1 \times C_2N_2\right)$ and the observed LR image as $y = \left[y_{k,1}, y_{k,2}, \ldots, y_{k,M}\right]^T$. If the parameters $C_1\left(0 < C_1 < 1\right)$ and $C_2\left(0 < C_2 < 1\right)$ respectively represent the down-sampling factors in the problem model for the horizontal and vertical directions, it comes out $M = N_1 \times N_2$. In the situation of $k > 1$, the problem model is considered to be based on multiple-images. While in the situation of $k = 1$, it is considered as the SR image reconstruction model based on single image which is expressed as Figure 4.5.

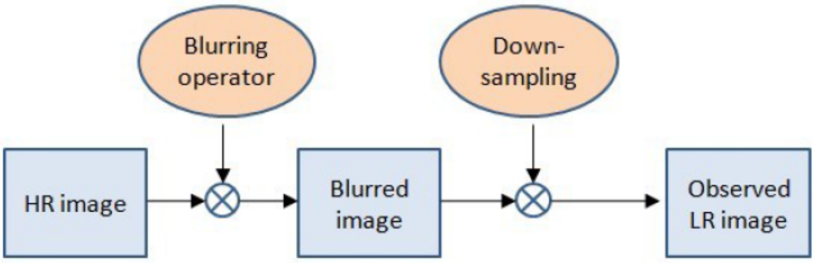


Figure 4.5 Problem model relating LR image to HR image.

In such case, the problem model can be shown as

$$y = DGf \, . \tag{4.12}$$

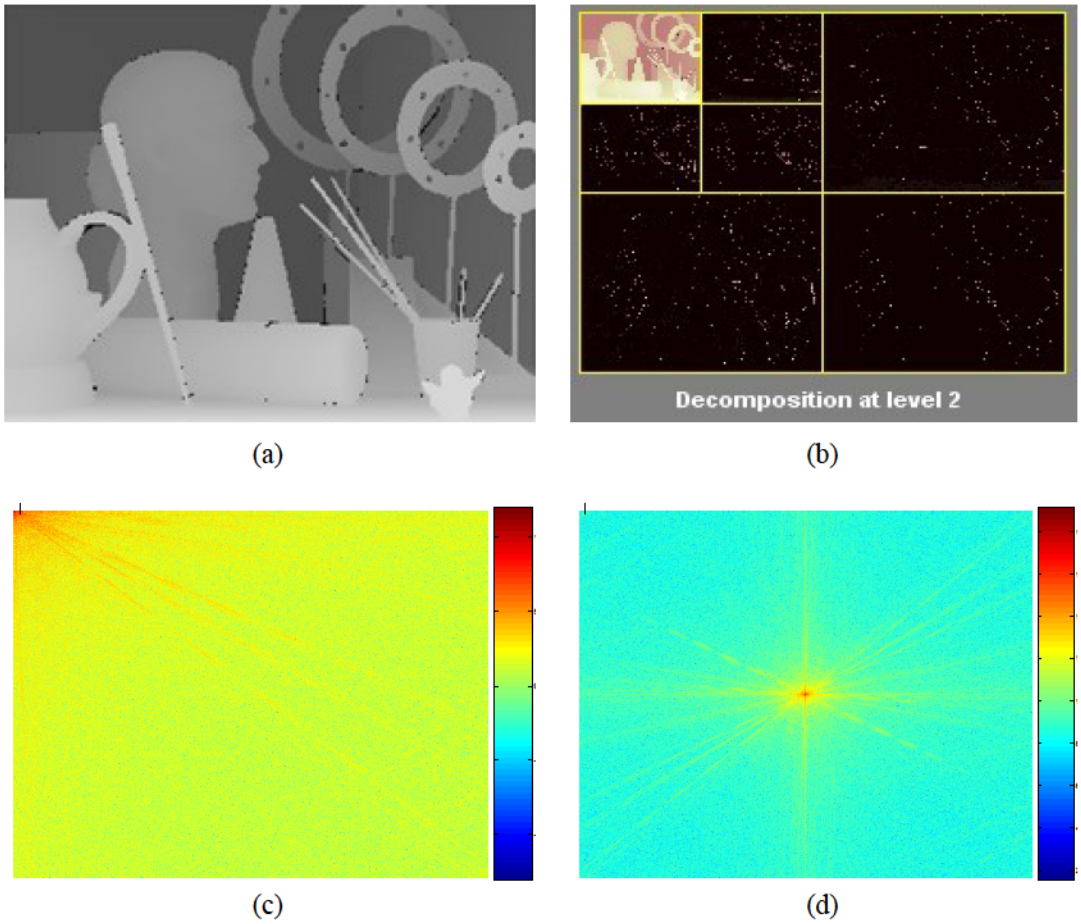
The goal of SR approaches is to reconstruct the unknown, desired HR image from the known, observed LR images. Mathematically speaking, the SR issue is an inverse problem. As it is obviously observed, the size of LR image M is less than that of HR image N, i.e., $M \ll N$. This SR issue is also an extremely ill-posed problem since for a given LR image y, there exist infinitely many HR images f meet the reconstruction constraint. In order to solve the ill-posed inverse problem, one needs to have some prior knowledge on the kind of typical images of interest. The prior information should help reconstruct the missing information.

Sparsity prior:

Sparsifying Transform: Natural, real-life images are known to be sparsely represented in the transform domain (e.g. wavelet, DCT and FFT domains). The wavelet transform is a multi-scale representation of the image which is usually used in Joint Photographic Experts Group (JPEG)-2000 image compression standard. Coarse-scale and fine-scale wavelet coefficients respectively represent the low resolution image components and the high resolution component. Each wavelet coefficient carries both spatial position and spatial frequency information at the same time. While the DCT and FFT are usually used for the JPEG image compression standard and Moving Picture Experts Group (MPEG) video compression, and it is of course can be utilized to sparsely represent images.

The sparsity of depth image: The transform sparsity of images can be demonstrated by applying a sparsifying transform to an image and reconstructing an approximation to the image from a subset of the largest transform coefficients. The sparsity of the image is the percentage of transform coefficients sufficient for evaluating reconstructions of image qualitatively and quantitatively. For purpose of illustration, the depth image is respectively transformed by the sparsifying transform of wavelet, DCT and FFT, respectively, as shown in Figure 4.6. As it can be observed that a depth image is highly sparse when represented by wavelet transform, DCT and FFT. This is not an unreasonable since the depth map is a digital image by coding the distance between the sensor and the object point of the scene.

For instance, the wavelet transform consists of recursively dividing the image into its low-frequency and high-frequency components. The lowest frequency components provide a coarse scale approximation of the image, while the higher frequency components fill in the detail. As most coefficients in the wavelet transform of the depth image are zero or very small, one can obtain a good approximation of the image via its K-sparse representation by setting the very small coefficients to zero.



(a) (b) (c) (d)

Figure 4.6 Sparse representation of a depth image via sparsifying transforms. (a) Original image. (b) Wavelet representation. (c) DCT representation. (d) FFT representation.

With the fact that the image SR reconstruction is an ill-posed inverse problem and the image sparsity prior in place, we put this problem model into the framework of compressive sensing as the recently emerging inversion theory which provides a great tool for solving the ill-posed inverse problem.

4.3.2 Compressive sensing model

The theory of compressive sensing states that, a signal can be exactly recovered from a small number of random linear measurements if it is sparse in some basis through non-linear optimization [1][11].

Signal sparsity: In a typical framework of compressive sensing, signal sparsity is significantly emphasized as the strong prior knowledge and vital prerequisite. That is, a signal vector $f \in \mathbb{R}^N$ can be sparsely represented by the form $f = \Psi x$, where $\Psi \in \mathbb{R}^{N \times N}$ denotes an orthonormal basis Ψ, and $x \in \mathbb{R}^N$ satisfies $\|x\|_0 = K$, $\|\cdot\|_0$ is l_0 norm, which means the number of its non-zero values, and $K \ll N$.

Random measurements: Due to the sparsity of f is relative to the orthonormal basis Ψ, it is not necessary to sample all N values of f. Instead, the CS theory establishes that f can be recovered from a small number of projections onto an incoherent set of measurement observations [1][11]. To measure y, we compute $M \ll N$ linear projections of f via the matrix-vector multiplication as Eq. 4.13.

$$y = \Phi f, \tag{4.13}$$

where $\Phi \in \mathbb{R}^{M \times N}$ is the sensing matrix or the measurement matrix. Further, the compressive sensing model is expressed as

$$y = \Phi f = \Phi \Psi x. \tag{4.14}$$

As aforementioned in Chapter 2, the sensing matrix Φ that fulfills the RIP includes i.i.d. Gaussian random matrices, Bernoulli matrices, and partial Fourier matrices [13]. An alternative approach to stability is to ensure that the sensing matrix is incoherent with the sparsifying basis Ψ.

Signal Reconstruction algorithm:

Besides the request of the conditions that a signal can be sparsely represented in a transform basis and the sensing matrix satisfies the RIP, the framework of compressive sensing depends as well as mainly on the reconstruction algorithms in terms of accuracy and speed. The incomplete list of reconstruction algorithms collection is shown in Section 2.1.3.

4.3.3 Super-resolution modeling via compressive sensing

By comparing the problem model with the CS model, we observe that the similarity between them is the inverse and ill-posed problem (i.e., both of them attempt to estimate the desired high-dimensional signal from the observed low-dimensional signal). Initially, it seems not possible to solve since the M samples of y yield a $N-M$ dimensional subspace of possible solutions for the original f that would match the given observations. However, the sparsity and the incoherence play a crucial role in solving ill-defined inverse problem. With a sensing matrix with random coefficients, Candes and Tao [9][12][13] and Donoho [11] proved that the inverse ill-posed problem can be solved for signals having a sufficiently sparse representation in some basis. The remarkable result offers a crack to the non-possibility and opens a door to recover high-resolution signals from a few randomized linear measurements. It is not an unreasonable assumption since the depth map is a digital image by coding the distance between the sensor and the object point of the scene and a digital image usually can be sparsely represented in a transform basis. We consider using the wavelet basis as the sparsifying basis Ψ to highly sparsely present the depth map since wavelet basis is very good at sparsely representing the image data.

Sparse representations

Signals or images carry overwhelming amounts of data in which relevant information is often more difficult to find than a needle in a haystack. Processing becomes faster and simpler in a sparse representation where few coefficients reveal the information we are looking for. Such representations can be constructed by decomposing signals or images over elementary waveforms chosen in a family called a *dictionary*. The discovery of wavelet orthogonal bases has opened the door to a huge jungle of new transforms. Adapting sparse representations to image properties is therefore a necessary survival strategy.

An orthogonal basis is a dictionary of minimum size that can yield a sparse representation if designed to concentrate the signal energy over a set of few vectors. This set gives a geometric signal description. Efficient signal or image compression is then implemented with diagonal operators computed with fast algorithms. Typically Fourier and Wavelet bases are usually used as transform bases for sparsifying signals and images. They decompose signals over oscillatory waveforms that reveal many signal properties and provide a way to sparse representations. Fourier and wavelet transforms illustrate the strong connection between well-structured mathematical tools and fast algorithms.

In the work, we use wavelet basis to sparsify the images as wavelet transforms have advantages over traditional Fourier transforms, especially because local features can be described better with wavelets that have local extent. A wavelet is a mathematical function used to divide a given function into different frequency components. A wavelet transform is the representation of a function by wavelets, which represent scaled and translated copies of a finite-length or fast-decaying oscillating waveform (known as the "mother wavelet"). Wavelet analysis represents a windowing technique with variable-sized regions and allows the use of long time intervals where more precise low-frequency information is needed, and shorter regions where high-frequency information is necessary. Wavelet bases reveal the signal regularity through the amplitude of coefficients, and their structure leads to a fast computational algorithm. Wavelet bases are well localized and few coefficients are needed to represent local transient structures. A wavelet basis defines a sparse representation of piecewise regular signals, which may include transients and singularities. In images, the large wavelet coefficients are located in the neighborhood of edges and irregular textures.

Inspired by the original ideas developed in computer vision by Burt and Adelson [102] to analyze images at several resolutions, Meyer and Mallat established the systematic theory for constructing orthonormal wavelet bases through the elaboration of multi-resolution signal approximations [105].

Wavelet for images

The principle of the wavelet decomposition is to transform the original raw image into several components with single low-resolution component corresponding to low frequencies or smooth parts of an image called approximation and the other components which represent the high frequencies called details, as shown in Figure 4.7. After applying bi-orthogonal low-pass wavelet in horizontal and vertical direction, and sub-sampling each image by a factor of two for each dimension, the approximation component can be extracted. The details are obtained by applying low-pass filter in one direction and a high-pass in the other direction, or alternatively a high-pass in both the directions. The noise is mainly presented in the details components. A higher level of decomposition is implemented by repeating the same operations on the approximation [53].

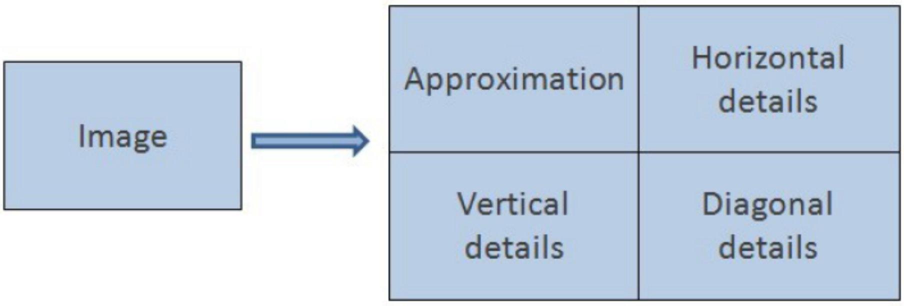


Figure 4.7 Wavelet decomposition of a 2D image.

A 2D image can be considered to be a matrix with *a* rows and *b* columns. The horizontal data is filtered at every level of decomposition, and then the approximation and details that produced from this are filtered on columns. Thus, at each level the four sub-images that are the approximation, the vertical detail, the horizontal detail and the diagonal detail are obtained. Likewise, the next level decomposition can be made via decomposing the approximation sub-image. The multilevel decomposition of an image is as shown in Figure 4.8.

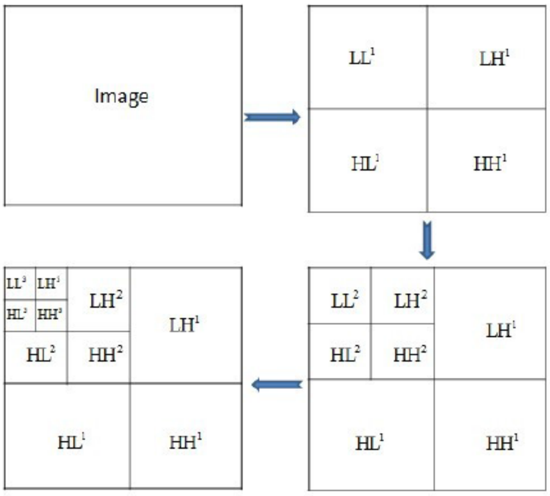
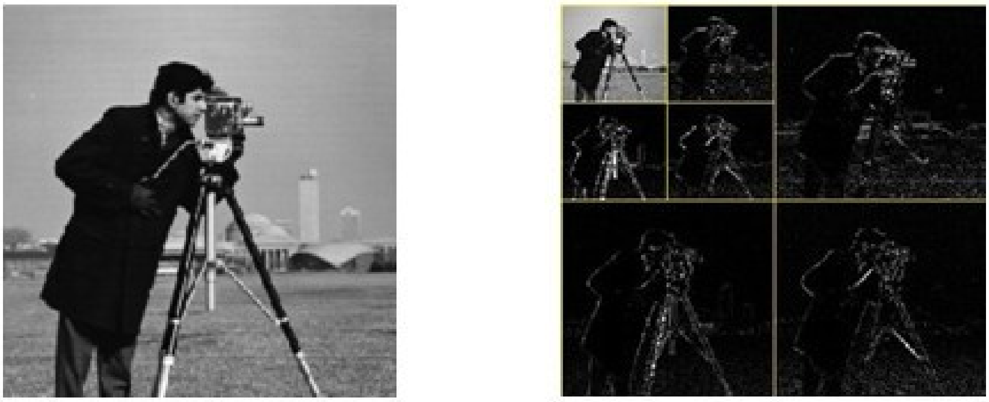


Figure 4.8 Multilevel (3 level) wavelet decomposition of an image.

Wavelet orthonormal bases of images can be constructed from wavelet orthonormal bases of one-dimensional signals. Three mother wavelets $\psi^1(x), \psi^2(x)$, and $\psi^3(x)$, with $x = (x_1, x_2) \in \mathbb{R}^2$, are dilated by 2^j and translated by $2^j n$ with $n = (n_1, n_2) \in \mathbb{Z}^2$. This yields an orthonormal basis of the space $L^2(\mathbb{R}^2)$ of finite energy functions $f(x) = f(x_1, x_2)$:

$$\left\{ \psi_{j,n}^k(x) = \frac{1}{2^j} \psi^k \left(\frac{x - 2^j n}{2^j} \right) \right\}_{j \in \mathbb{Z}, n \in \mathbb{Z}^2, 1 \le k \le 3} . \tag{4.15}$$

The support of a wavelet $\psi_{j,n}^k$ is a square of width proportional to the scale 2^j. 2D wavelet bases are discretized to define orthonormal bases of images including N pixels. Like in one dimension, a wavelet coefficient $\langle f, \psi_{j,n}^k \rangle$ has a small amplitude if $f(x)$ is regular over the support of $\psi_{j,n}^k$. It has a large amplitude near sharp transitions such as edges. The following figure is the array of N wavelet coefficients, as shown in Figure 4.9. Each direction k and scale 2^j corresponds to a sub-image, which shows in black the position of the largest coefficients above a threshold: $\left| \langle f, \psi_{j,n}^k \rangle \right| \ge T$.



(a)

(b)

Figure 4.9 (a) Discrete image. (b) Array of orthogonal wavelet coefficients.

Wavelets Overview

The essence behind wavelets is to analyze arbitrary signals according to its scales in frequency domain. Hence, it is a type of multi-resolution analysis. Wavelets are that the functions which are defined over a finite interval. They are obtained from a single prototype wavelet called mother wavelet by dilation and translation at different positions and on different scales. An arbitrary signal can be represented as a linear combination of such wavelet, or basis functions.

Haar wavelets

The Haar wavelet is the basis of the simplest wavelet transform. And it is the only one that has an explicit expression in discrete form and the only symmetric wavelet in the Daubechies family as well. Haar wavelets are related to a mathematical operation called Haar transform, which serves as a prototype for all other wavelet transforms.

The Haar basis is obtained with a multi-resolution of piecewise constant functions. The scaling function is $\tau = 1_{[0,1]}$. The filter $h[n]$ has two non-zero coefficients equal to $2^{-1/2}$ at $n=0$ and $n=1$. Haar constructed a piecewise constant function which is the so-called one-dimensional Haar wavelet:

$$\psi(t) = \begin{cases} 1 & \text{if} \quad 0 \le t \le 1/2 \\ -1 & \text{if} \quad 1/2 \le t \le 1 \\ 0 & \quad \textit{otherwise} \end{cases} . \tag{4.16}$$

Haar wavelet transform only provides non-redundant representation of the signal since the Haar basis is an orthogonal basis. The 2D Haar wavelet is an extension of one-dimensional Haar wavelet. For image processing, the standard 2D Haar wavelet transform can be implemented as one-dimensional Haar wavelet transform applied on rows of image followed by another one-dimensional Haar transform applied on the columns of the transformed image [102].

Daubechies wavelets

Daubechies wavelets are a family of scale functions that are orthogonal and have finite vanishing moments (i.e., compact support) [90]. This property insures that the number of non-zero coefficients in the associated filter is finite. The Daubechies wavelet transforms are defined in the same way as the Haar wavelet transform by computing the running averages and differences via scalar products with scaling signals and wavelets.

Besides Daubechies wavelets, there exist some other mother wavelets families such as Symlet, Coiflet, Biortogonal and Reverse biorthogonal wavelet are also used for many applications.

Sensing matrix

As discussed before, the problem of super-resolution reconstruction is formulated as the model expressed as

$$y = DGf \tag{4.17}$$

and the compressive sensing model as

$$y = \Phi f \, . \tag{4.18}$$

If we suppose to put the problem model into the compressive sensing model, it has to make $\Phi = DG$, which means the sensing matrix is constructed by the combined operators of down-sampling and blurring. The RIP indicates that if every set of the sensing matrix columns with cardinality less than the sparsity of the signal of interest is approximately orthogonal, the signal can be exactly reconstructed with high probability. The entries of such sensing matrix can be taken from Gaussian distribution, symmetric Bernoulli distribution, etc. The Gaussian smoothing operator is usually used to “blur” images and remove detail and noise. The blur version of the desired HR image can be described as:

$$f_{blur} = Gf \, . \tag{4.19}$$

We consider the Gaussian filter as a multiplication of a Gaussian function in the Fourier domain, as expressed in Eq. 4.20.

$$G = F^{-1} G_* F \quad , \tag{4.20}$$

where F denotes the 2D Fourier transform matrix and F^{-1} is the inverse Fourier transform matrix, G_* denotes the diagonal matrix whose diagonal elements correspond to the Gaussian function and elsewhere is zero. The sensing matrix is then expressed as

$$\Phi = DF^{-1} G_* F \quad . \tag{4.21}$$

And the super-resolution problem can be modeled in the framework of compressive sensing as the following:

$$y = DF^{-1} G_* F \Psi x \quad . \tag{4.22}$$

In this formulation, y is regarded as known LR image and f as unknown HR image.

Signal reconstruction

With this formulation in hand, given a LR depth image y, we design a sensing matrix which consists of random down-sampling operator and Gaussian filter to meet the RIP and apply signal reconstruction algorithms to solve the optimization problem:

$$\hat{x} = \arg\min \|x\|_1 \quad s.t. \quad y = DF^{-1} G_* F \Psi x \quad . \tag{4.23}$$

Then the HR depth image can be reconstructed from the sparse representation $\hat{x}$, as expressed in Eq. 4.24.

$$f = \Psi \hat{x} \quad . \tag{4.24}$$

4.4 **Experimental results**

In this section, we demonstrate the experimental results of two image samples to evaluate the performance of the proposed method. For the first test sample, we first directly apply the Gaussian low-pass filter operator and decimation operator to the original HR image in order to obtain the LR image. We adopt the proposed approach on the observed LR image to reconstruct the desired HR image. The reconstructed HR image is compared with the original HR image using the compared methods include perceptual quality evaluation and quantitative evaluation. For the second test sample, we would like to directly apply the proposed method to obtain the HR depth map, since only with the LR depth image as the measurement observations in hand instead of HR depth image.

4.4.1 Simulated results

Since the real depth map suffers from both lower resolution and high noise, in the first test sample we use a standard grayscale image which often is used in image processing community to evaluate the performance of the proposed method. We first use the wavelet transform with 'haar' at level 2 to decompose the original image, as shown in Fig.4.10 (b). The original image shown in Fig.4.10 (a) exhibits the highly sparsity under the wavelet basis. With the sparsity property in hand, we then blur it with the Gaussian low-pass filter and obtain a sub-sample version with down-sample mask, as shown in Fig.4.10 (c).

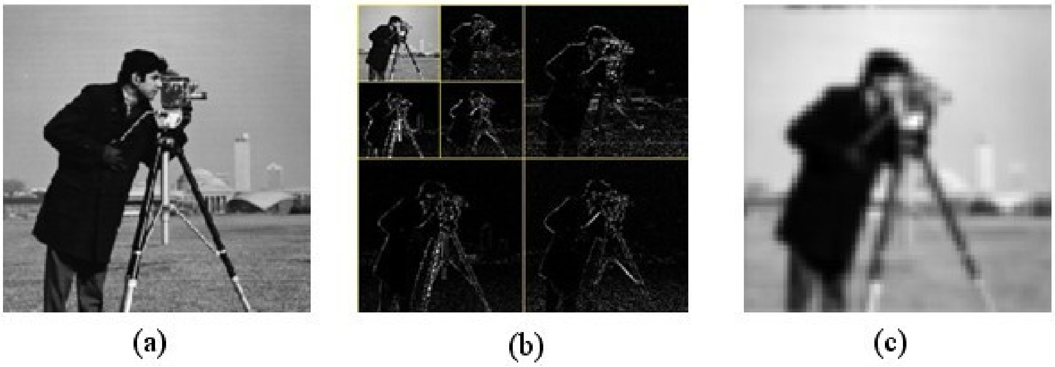


Figure 4.10 (a) The original HR image. (b) Wavelet decomposition. (c) The LR image.

The images are up-scaled with various algorithms for comparison purpose. As shown in Fig.4.11 we compare the results of our method with the standard approaches: bilinear interpolation and bicubic interpolation.

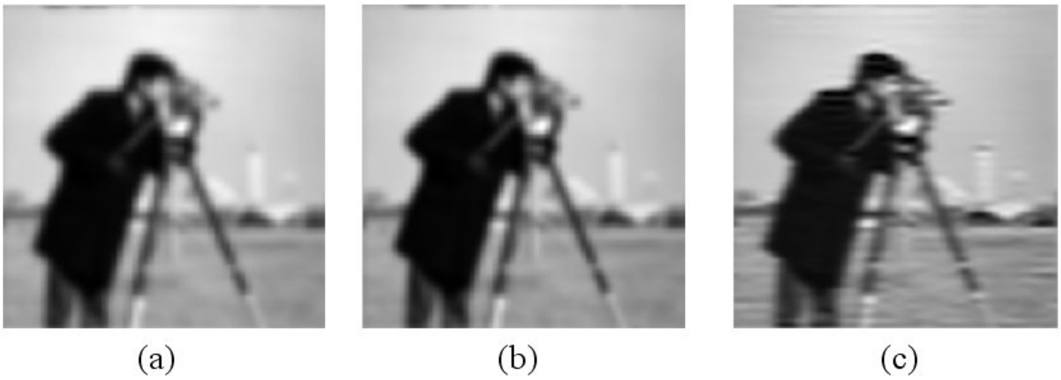


Figure 4.11 (a) SR result via 'bilinear' interpolation. (b) SR results via 'bicubic' interpolation. (c) SR result via the proposed method.

From perceptual point of view, the result of our method performs better than the other approaches with regard to sharper details. For example, one can observe the cameraman's eye in the image reconstructed by the proposed approach. In the meanwhile, we evaluate the results with Peak Signal-to-Noise Ratio (PSNR), the most widely used objective image quality metric. However, an interesting result is that the PSNR value from the proposed method is lower than the other methods. As we have known, the PSNR value does not perfectly correlate with a perceived visual quality due to the non-linear behavior of the human visual system.

4.4.2 Real depth image

The depth map (i.e., Fig.4.12 (a)) used in this work was 204×204 pixels in size with a bit depth of 8 bits. It was captured from the PMD camera within the MultiCam monitoring a natural scene with a man walked in the field of view of the camera. However, a big difference from the first test sample is that we directly apply the proposed method to the depth map instead of the down-sample version of HR image due to the fact that only LR depth map is available. We analyze the sparsity of the original LR depth map in wavelet domain. As we can observe from Fig. 4.12 (b), it can be sparsely represented in wavelet transform at level 2. Likewise, it is not nonreasonable that the desired HR depth map is also sparse in the wavelet domain. With the precondition in hand, we therefore can invoke the super-resolution method via compressive sensing framework: first apply the Gaussian low-pass filter and then point down-sample mask to the desired HR depth map. With y as the given LR depth map, we compute the sparse coefficients under the wavelet basis Ψ and afterwards reconstruct the desired HR depth image using the equation of $f = \Psi x$.

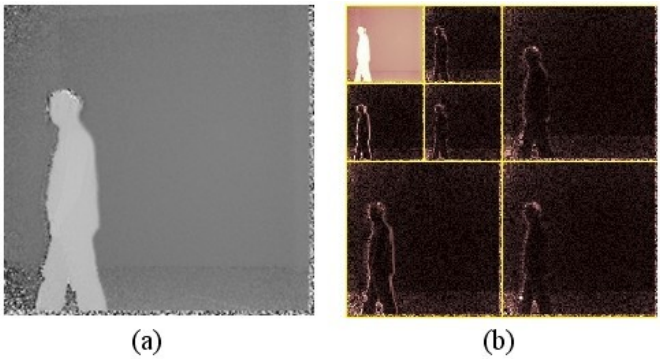


Figure 4.12 (a) Original LR depth map. (b) Wavelet transforms.

For the clarity, we use a part of the results from interpolation and the proposed methods. The part of the original LR depth map is shown in Fig.4.13 (a). The results are shown in Fig.4.13 (b) and (c), respectively. As it can be seen, the proposed method produces the smoother and clearer edge than the other methods. And meanwhile we should also note that the proposed method not only enhances the edge, but also implements the function of noise removal to some extent. Therefore, the proposed method produces a better performance.

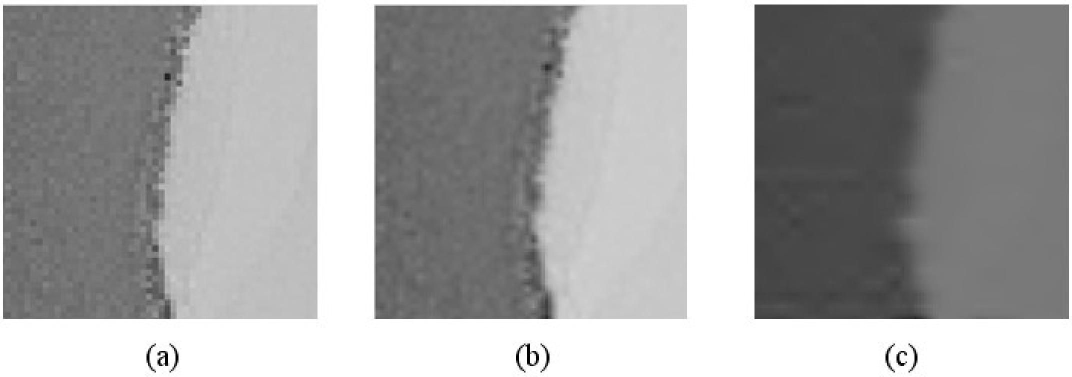


Figure 4.13 (a) Part of Original LR depth map. (b) SR result via 'bilinear' interpolation. (c) SR result via the proposed method.

4.5 Chapter summary

We discuss the main range measurement techniques and advantages and disadvantages including structured light, stereoscopy, laser pulse range finder and Time of Flight. We also study the problem model of depth image super-resolution and compressive sensing theory model and develop the super-resolution signal model via the framework of compressive sensing by the comparison and observation of the both models. The depth map is sparsely represented under the wavelet basis. As shown in the results, the 'haar' wavelets at level 2 we used for compression worked well in the standard gray-level image, while the 'db' wavelets at level 2 worked well in the depth image. We suppose that there might be better-suited wavelet basis for the sparsifying basis. The sensing matrix is designed with the multiplication of a point down-sample mask and a Gaussian low-pass filter and it is sensitive to the choice of the sparsifying basis according to the RIP, which means it seems there still exists better sparsifying basis such as complex wavelets for sparse representation. In the meanwhile, the random point down-sample mask can be used to improve the effectiveness of the proposed method. In addition, the reconstruction algorithms also play an important role for the result. The other recover algorithms would be interesting to explore.

5 Sensing change in multimodal images

Sensing regions of change in multiple images of the same scene taken at different times is of widespread interest and critical element due to a large number of applications in diverse disciplines. Some of important applications include video surveillance, monitoring, remote sensing, civil infrastructure and so forth. According to various applications, the researchers pay more attention to the processing steps and core algorithms.

In this chapter, an overview of previous work concerning change detection in images will be given. Then we study and discuss the proposed sparse feature based approach, followed by experimental results.

5.1 Multimodal image

5.1.1 2D/3D vision system

In order to take the advantage of different modal sensors, the combinational use of multimodal sensors has been becoming a tendency in the recent research work. Many researchers have worked on the related field and proposed some important sensor combinations. Generally speaking, a 2D/3D vision system denotes the combinational use of 2D vision sensor and 3D range sensor. A brief overview of combinational use of them is given as below.

- Combinational use of ToF camera with 2D color camera.

Ghobadi et al. in [38] used the combined 2D camera and 3D ToF camera for analysis of the personnel safety in man-machine cooperation and in [37] they utilized 2D/3D images for hand segmentation. Van den Bergh et al. [71] combined RGB and ToF cameras for real-time 3D hand gesture.

- Combinational use of ToF camera and stereo vision systems

Netramai et al.[73] combined PMD and stereo camera for motion estimation of a mobile robot. Hahne et al. [74] combined ToF camera and on-demand stereo for depth imaging.

- Combinational use of laser rangefinder with a standard 2D camera system.

Klimentjew et al. [66] used a perception system which consists of a camera and a 3D laser range finder do multi sensor fusion for object recognition. Joung et al. [69] proposed a system which reconstructed the environment with both color from 2D camera and 3D information from a laser range finder.

- Combinational use of a stereo vision system with a laser range finder.

Aliakbarpour et al. [68] presented a new and efficient method to perform the extrinsic calibration between a 3D laser range finder and a stereo camera with the help of inertial data. Monteiro et al. [70] exploited the combined laser rang finder and vision to track and classify the dynamic obstacles.

5.1.2 MultiCam

Among many approaches to combine multi-modal sensors, the use of combination of a standard color imager and a ToF sensor in a monocular setup called MultiCam [5], as shown in Figure 5.1, has been developed in the Center for Sensor System (ZESS) and becomes one of the promising techniques in 2D/3D imaging system.

The so-called MultiCam consists of two imaging sensors: a conventional standard CMOS sensor with Video Graphics Array (VGA) resolution and a PMD ToF sensor, a beam splitter, a near-infrared lighting system, Field Programmable Gate Array (FPGA) based processing unit and USB 2.0 or Gigabit Ethernet communication interface. A previous version uses the 3K PMD chip with resolution of 64×48 pixels. The newest version uses the 41K PMD chip with resolution of 204×204 pixels. While the 3D sensor needs to acquire the modulated near-infrared light (in our case 870 nm) back from the scene, the 2D sensor is used to capture the images in the visible spectrum (approximately 400 nm to 800 nm). To do this, a dichroic beam splitter (Dichroic beam splitters are used to combine or separate beams of two different wavelengths) behind the lens has been used which divides the acquired light into two spectral ranges: the visible light which is forwarded to the 2D sensor and the near-infrared spectrum which is directed to the 3D sensor.

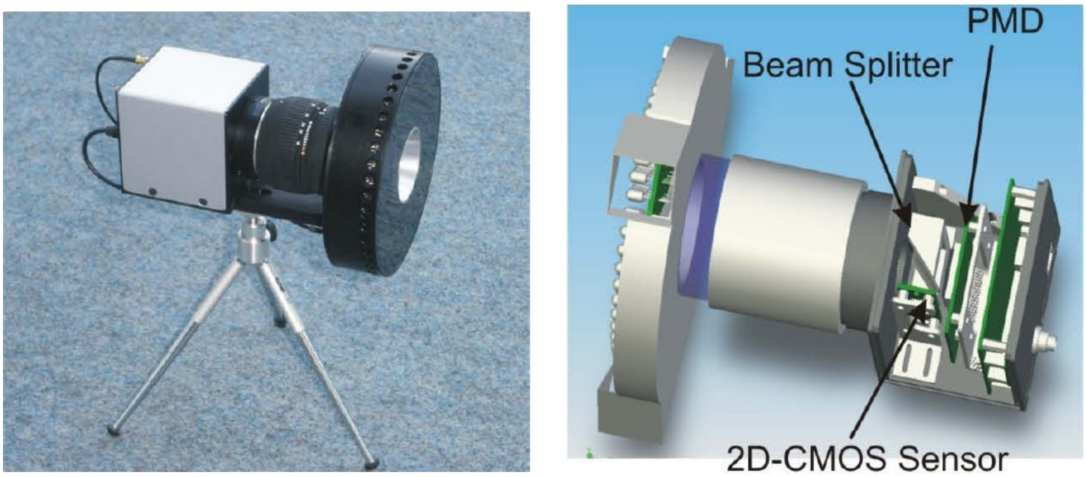
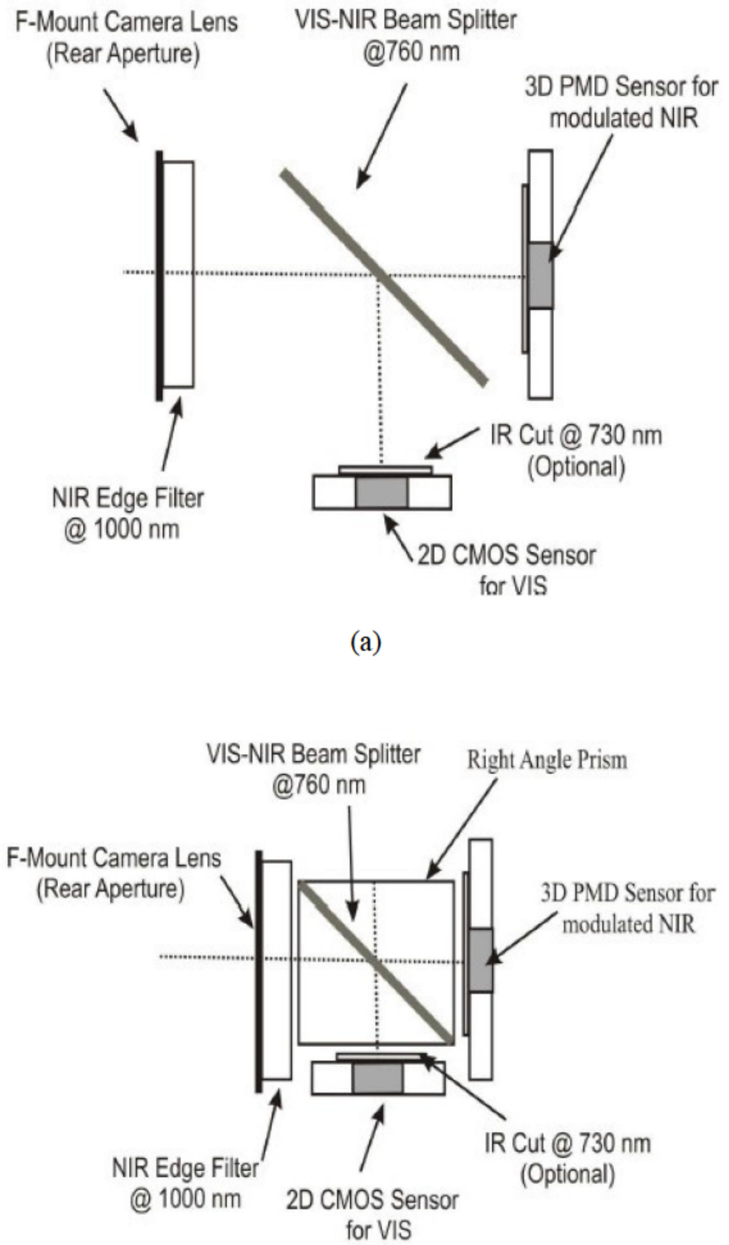


Figure 5.1 The MultiCam 2D/3D vision system

The MultiCam with two different optical designs is available: F-mount and C-mount which are illustrated in Fig. 5.2. The F-mount optical design has a simple setup due to its large flange focal distance (The flange focal distance is 46.500 *mm* for a F-mount, whereas it is 17.526 *mm* for a C-mount lens.) which makes positioning of the chips as well as their adjustment in the setup simple. In this case, a beam splitter which is a commercial cold mirror is fixed at the angle of 45° with the rear surface being anti-reflection-coated for the near-infrared spectrum. In fact, such a coating is the crucial part of optical design. However, F-mount is not suitable for $1/2''$ chip formats like PMD sensor because the large focal distance of F-mount with a small chip size of $1/2''$ yield a narrow angle of view which is not suitable for some applications. On the other hand, C-mount lenses are good options for the chips with $1/2''$ format. However, in a C-mount design the flange focal distance is shorter than in an F-mount which consequently makes the mechanical design and adjustment of the sensors in the setup more complicated. One solution to this problem, which is used in the design of the C-mount MultiCam, is to use a prism beam splitter, as illustrated in Figure 5.2 and Figure 5.3. In fact, in this case the beam splitter is placed between two prisms made out of glass. As the glass has a higher refractive index, the optical path length gets bigger which consequently increases the focal length. In other words, by using the prisms made of glass one can lengthen the distance between the lens and the sensors which makes the arranging as well as adjusting the chips in the optical setup easier.



(a)

(b)

Figure 5.2 (a) F-Mount. (b) C-Mount.

5.2 Related work

Due to the fact that change detection in multimodal image plays a crucial role in various application areas, there is a rich research body in the domain related to change detection using multimodal image. And a number of researchers have made much effort to obtain good results by applying a variety of methods.

The application that the region of interest contains only hand is extracted and used for hand tracking from 3D depth image as well as gesture recognition was presented in [37]. Plagemann et al. [94] used a novel interest point descriptor together with a boosted patch classifier to localize body parts using range data. Chen et al. [107] applied a region growing technique to segment the hand region on depth images. Then a mean-shift based algorithm accurately locates the hand center in the segmented hand region. Ghobadi et al. [38] used the combination of edge detection and an unsupervised clustering technique for foreground segmentation. And a rather simple approach to extract foreground from 2D/3D videos which is based on region growing and refrains from modeling the background was evaluated in the work of Bianchi et al. [41]. Van den Bergh et al. [71] used a simple threshold technique to separate the background and the foreground, given a correct mapping from the depth data to the RGB image. Schoenberg et al. [108] proposed that the textured 3D dense point cloud is then segmented based on evidence of a boundary between regions of the textured point cloud after fusion of camera image and laser range data using Markov Random Field to estimate a 3D point corresponding to each image pixel. In the work of Leens et al. [39] the pixel-based background modeling method, called ViBe, was separately applied to the RGB channels of color image and the three channels of PMD camera. The resulting foreground masks were combined via binary image operations. In [97] the ability of bilateral filtering to deal with geometric objects was demonstrated. A more elaborate method of fusing color and depth was bilateral filtering, which has been used in [50], where the preliminary foreground was produced by a dividing plane in space and a bilateral filter was applied to gain the final result. Harville et al. [42] proposed a method for modeling the background that uses per-pixel, time-adaptive, Gaussian mixtures in the combined input space of depth and luminance-invariant color. They improved such combination by introducing the ideas of (1) modulating the background model learning rate based on scene activity, and (2) making color-based segmentation criteria dependent on depth observations. The input to the algorithm is a time series of spatially registered, time-synchronized pairs of color (YUV space) and depth images obtained by static cameras. Gordon et al. [94] modeled each pixel as an independent statistical process, recoding the (R, G, B, Z) observations at each pixel over a sequence of frames in a multidimensional histogram (depth and color information). Then they used a clustering method to fit the data with an approximation of a mixture of Gaussians. At each pixel, one of the clusters (Gaussians) has been selected as the background process, the others were considered to be caused by foreground processes. An approach working with one color image and multiple depth images is described in [96], where data fusion has been formulated in a statistical manner and modeled using Markov Random Fields on which an energy minimization method was applied.

5.3 Sensing change by sparse feature reconstruction

5.3.1 Image matrix decomposition

From image analysis point of view, each frame of a video sequence consists of two layers: background and foreground. We define the background as the static or approximately static region (more or less affected by ambient varying illumination) and the foreground as the region corresponding to the moving object. The image with almost stationary background and dynamic foreground can be considered as the samples of signals that change slowly in time with the sparse feature with arbitrary shape caused by dynamic foreground.

If the background and foreground of an image are denoted by B_i and F_i , the image I_i can be represented as:

$$I_i = B_i + F_i \, , \tag{5.1}$$

where i denotes a natural number. Assume given the images $I_i \in \mathbb{R}^{w \times h}, \quad i = 1, 2, \ldots, n$ which are taken from a scene at the different time, if we let

$vector:\mathbb{R}^{w\times h}\to\mathbb{R}^{N}\left(N=w\times h\right)$ denote the operator that stacks the image sequence into the column vector, then the image sequence matrix I can be formulated as

$$I=\left[vector\left(I_1\right)|\ldots|vector\left(I_n\right)\right]. \tag{5.2}$$

Suppose the matrix B formed by the background image sequences denoted as

$$B=\left[vector\left(B_1\right)|\ldots|vector\left(B_n\right)\right], \tag{5.3}$$

and the matrix F formed by the foreground image sequences from each images can be represented as

$$F=\left[vector\left(F_1\right)|\ldots|vector\left(F_n\right)\right]. \tag{5.4}$$

The image sequences matrix I can be formulated as the sum of background sequences matrix and foreground sequences matrix:

$$I=B+F. \tag{5.5}$$

However, it is a severely under-constrained problem. It is difficult to find B and F without any prior information. In the next sub-section we will analyze and find the prior for the problem.

5.3.2 Modeling change as sparse feature

Assume in ideal situation (i.e., there is no moving object in foreground) the background sequence matrix is a low-rank matrix with rank one. This leads us to build a low-rank constraint model for the background matrix B :

$$rank(B)=1. \tag{5.6}$$

However, in common situation background sequences taken from a static camera in a scene are linear correlated. The background sequence is therefore modeled by a low rank subspace that can gradually change over time and consequently the matrix B exhibits a low-rank structure. While the pixels in foreground sequences can be clustered into a gross sparse feature with arbitrarily large magnitude and most entries in the foreground matrix F will be zero, which means the foreground matrix is a sparse matrix.

The image sequence matrix is therefore considered to be the sum of a low-rank matrix and a sparse matrix. The problem can be formulated as exact recovery of the sparse matrix in order to reconstruct the sparse feature. We discuss more details about the reconstruction of sparse and low-rank matrix in the next sub-section.

5.3.3 Sparse feature reconstruction

To exactly recover the two components and furthermore reconstruct the sparse feature for the change region, it is suggested a conceptual solution: use l_0-norm to control the sparsity structure of the matrix F and seek the lowest rank B to encourage the low-rank structure. The Lagrangian reformulation of the problem is:

$$\min_{B,F} rank\left(B\right)+\lambda\left\|F\right\|_0 \quad s.t. \quad I=B+F \qquad (5.7)$$
,

where λ is a balance parameter that trades off the rank of the solution versus the sparsity of the matrix F. For appropriate balance parameter λ, however, this is a highly non-convex optimization problem. And it is known there is no efficient solution for it. By replacing the l_0-norm with the l_1-norm, and the rank with the nuclear norm, we relax the non-convex optimization problem and yield the following convex surrogate:

$$\min_{B,F} \left\|B\right\|_* + \lambda\left\|F\right\|_1 \quad s.t. \quad I=B+F \qquad (5.8)$$
.

The above problem can be treated as a general convex optimization problem and solved by any off-the-shelf interior point solver [20]. Although interior point solver has excellent convergence properties of normally taking very few iterations to converge, it is not quite scalable for large matrices problems.

Based on the recent work on recovery of corrupted low-rank matrix [36], we apply the method of augmented Lagrange multiplier to solve problem effectively.

The augmented Lagrange function is given by:

$$L_{\mu}\left(B,F,Y,\mu\right)=\|B\|_{*}+\lambda\|F\|_{1}+\left\langle Y,I-B-F\right\rangle+\frac{\mu}{2}\|I-B-F\|_{F}^{2}, \tag{5.9}$$

where $Y\in\mathbb{R}^{M\times N}$ is a Lagrange multiplier matrix, μ is a positive scalar, $\langle\cdot,\cdot\rangle$ denotes the matrix inner product, and $\|\cdot\|_F$ denotes the Frobenius norm. The augmented Lagrange multiplier method iteratively estimates both the Lagrange multiplier and the optimal solution by iteratively minimizing the augmented Lagrangian function:

$$\begin{aligned}&\left(B_{k+1},F_{k+1}\right)=\arg\min\nolimits_{B,F}L_{\mu k}\left(B,F,Y_{k}\right),\\&Y_{k+1}=Y_{k}+\mu_{k}\left(I-B_{k+1}-F_{k+1}\right)\end{aligned}. \tag{5.10}$$

The iteration definitely converges to the optimal solution of the problem in Eq. (5.8) [59]. However, it is very difficult to directly solve the first step in the above iteration in Eq. (5.10). Researchers commonly try to minimize the Lagrangian function approximately against the two unknown variables B and F at one time:

$$B_{k+1}=\arg\min\nolimits_{B}L_{\mu_k}\left(B,F_{k},Y_{k}\right), \tag{5.11}$$

$$F_{k+1}=\arg\min\nolimits_{F}L_{\mu_k}\left(B_{k+1},F,Y_{k}\right). \tag{5.12}$$

For the clearness, the complete procedure to solve the convex optimization based on the method of augmented Lagrange multiplier is summarized in Algorithm 5.1.

Implementation Issues

Due to the fact that the basic idea of the augmented Lagrange multiplier method is to search for the saddle point of the augmented Lagrange function, rather than directly solving the original constrained optimization problem. We vary the parameter μ starting from the initial value μ_0 and increase it monotonically with each iteration to speedup the convergence until it reaches sufficiently large, in the meanwhile the difference in the value of the cost function is small enough between two consecutive iterations when the iteration definitely converges to the optimal solution of the problem [59].

```
Algorithm 5.1: Optimization via Augmented Lagrange Multiplier Method
Input:  Observations Matrix $I \in \mathbb{R}^{M\times N}$, $\lambda$, $\sigma$, $\mu$.
while not converged do
    $(U,S,V) = svd\left(I - F_k + \frac{1}{\mu_k} Y_k\right);$
    $B_{k+1} = US_{\frac{1}{\mu_k}}[S]V^T;$
    $F_{k+1} = S_{\frac{\lambda}{\mu_k}}\left[I - B_{k+1} + \frac{1}{\mu_k} Y_k\right];$
    $Y_{k+1} = Y_k + \mu_k (I - B_{k+1} - F_{k+1});$
    $\mu_{k+1} = \rho\mu_k;$
    $k = k + 1;$

end while
Output: $B = B_k; F = F_k$.
```

where $svd(\cdot)$ denotes the singular value decomposition operator.

5.4 Experimental results

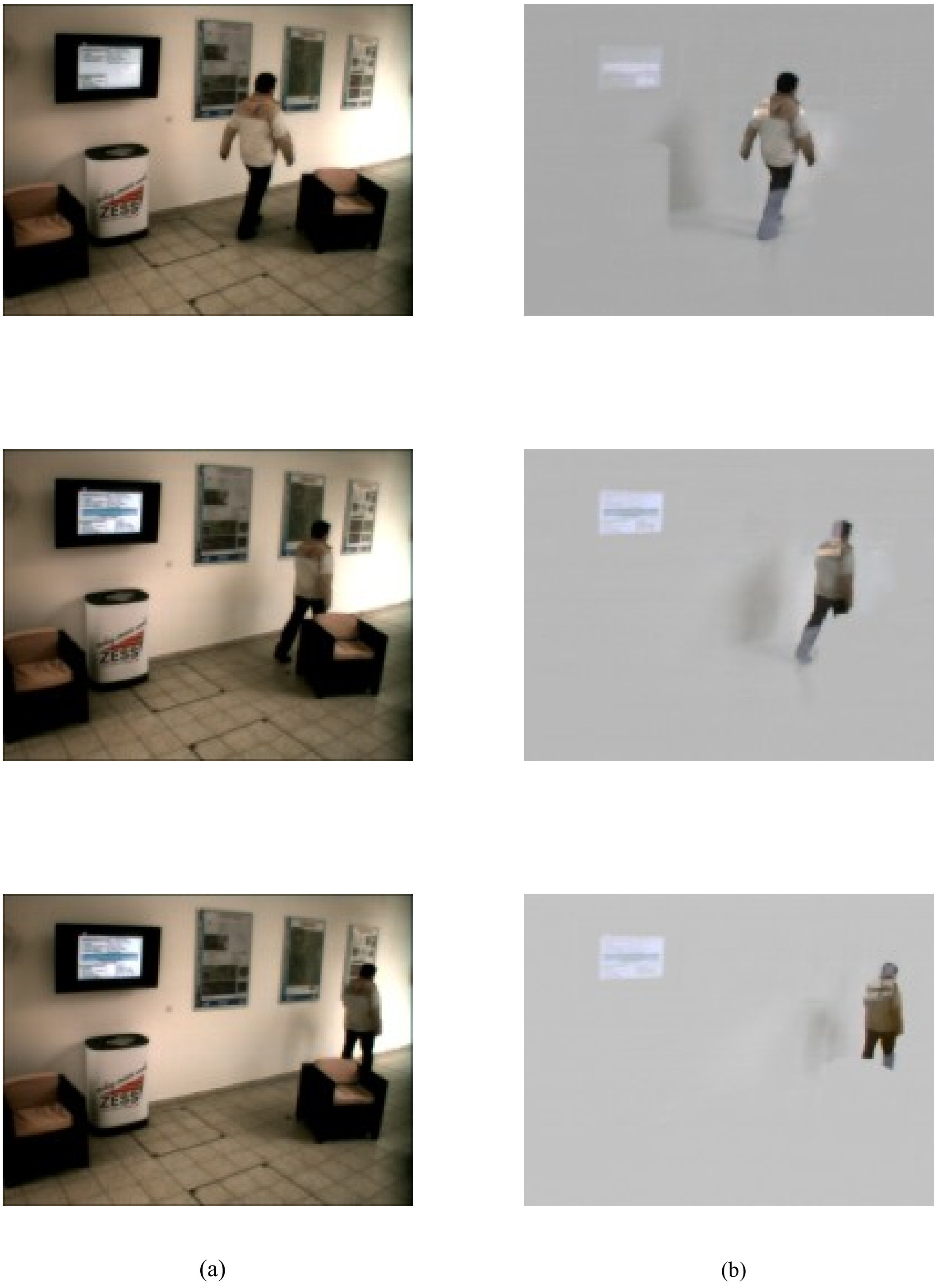
We verify the effectiveness and evaluate the performance of the proposed method with the experimental results on the two different image database recorded consists of 2D color image and corresponding depth map as well as modulation amplitude map that represent different situations for indoor video surveillance. Both of them for evaluation are recorded under normal lighting conditions and bad illumination, respectively. To demonstrate the primary performance of the proposed method, post-processing such as noise removal and connected component analysis are not introduced in the work.

Sequences 1

The sequence is an indoor scene consists of lobby where a person walks from very close to the MultiCam to far. There is a TV mounted on the wall whose screen is changing. The ambient lighting conditions are quite stable. And the walking man is well contrasted with the background. The random three frames from this sequence are presented in Figure 5.3 (a). Due to the memory limitation of the standard computer, the color image with 640×480 pixels is firstly down-sampled to 192×144 pixels using simple down-scale technique. To train these images, sequential 30 frames from the image sequences are used. And the balance parameter is set as $\lambda = 0.8$. It can be seen from the human detection results shown in Figure 5.3 (b) that the walking man is perfectly detected. In the meanwhile, however, the TV screen and the shadow cast by the walking man also are detected due to the fact that the two regions vary when the man is walking.

Since it is not possible to remove the varying background belongs to foreground, we directly apply the proposed method to the original depth image without deleting invalid measurement. Hereby, the balance parameter is fixed as $\lambda = 2$. The original depth images and human detection results from them are shown in Figure 5.4 (a) and Figure 5.4 (b), respectively. It can be seen the results in Figure 5.4 (b) contain much noise where belongs to invalid measurement in modulation amplitude map. If the active modulated infrared light returns from the human point in the scene cannot be sensed by PMD sensor, the pixel in the location should be zero in the corresponding modulation amplitude map. After removal of the invalid measurement for the original depth image shown in Figure 5.5 (a), it is apparent that the proposed method can remove the background details and exactly reconstruct the human, as shown in Figure 5.5 (b).

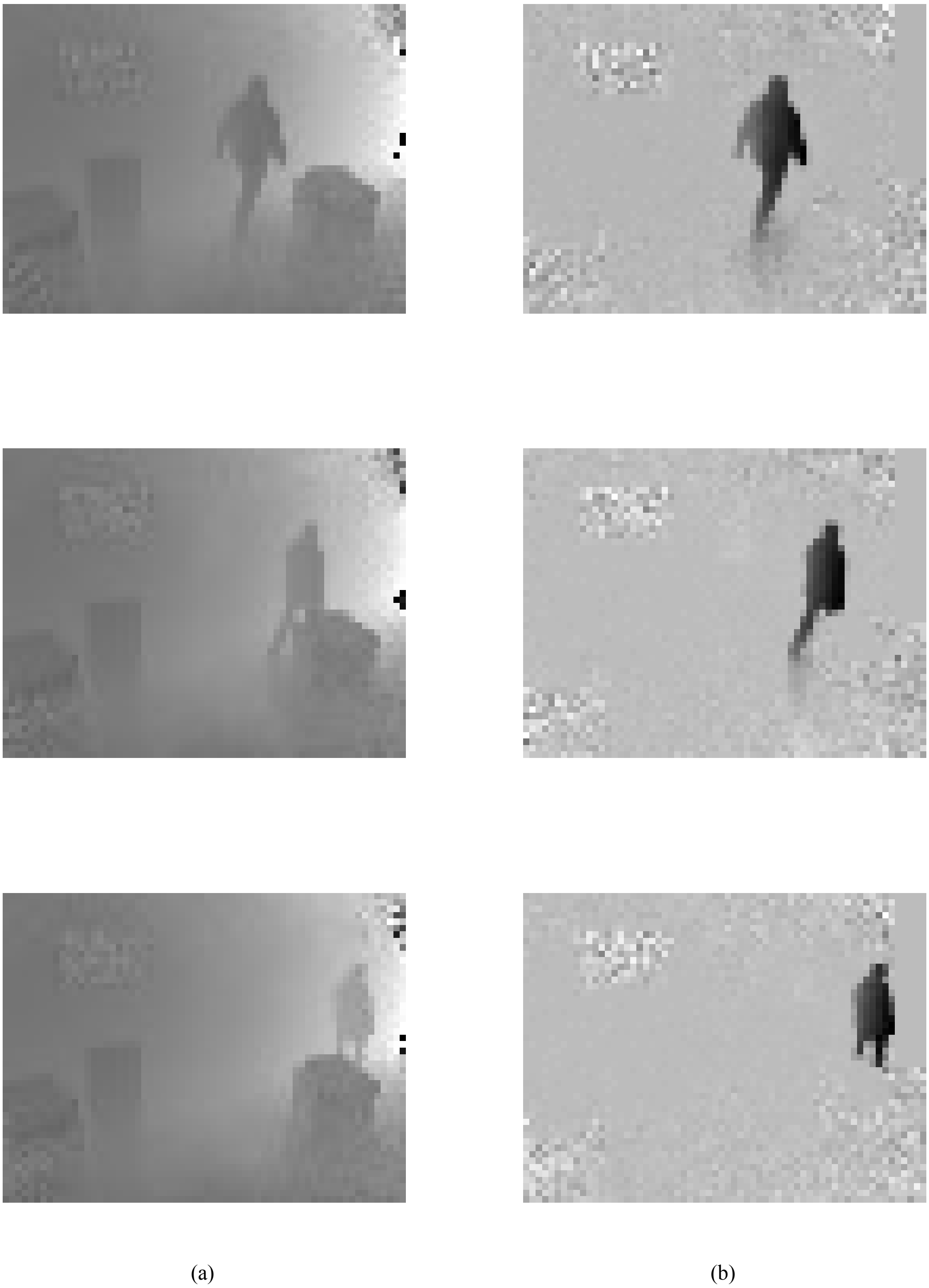
Simultaneously applying it to the modulation amplitude map as demonstrated in Figure 5.6 (a), we set g and obtain the results as shown in Figure 5.6 (b). Although the original depth map exhibits noisy representation, the extraction of distance information belongs to the region of moving human in depth map is dramatically improved after removal of invalid measurement according to the modulation amplitude map. The combination of the human detection results from 2D color images, 3D depth map and modulation amplitude map can supply big help for post-processing such as human tracking, location and so on that are not introduced in detail in the thesis.



(a)

(b)

Figure 5.3 (a) Color images. (b) Human detection results.



(a) (b)

Figure 5.4 (a) Original depth images. (b) Human detection results.

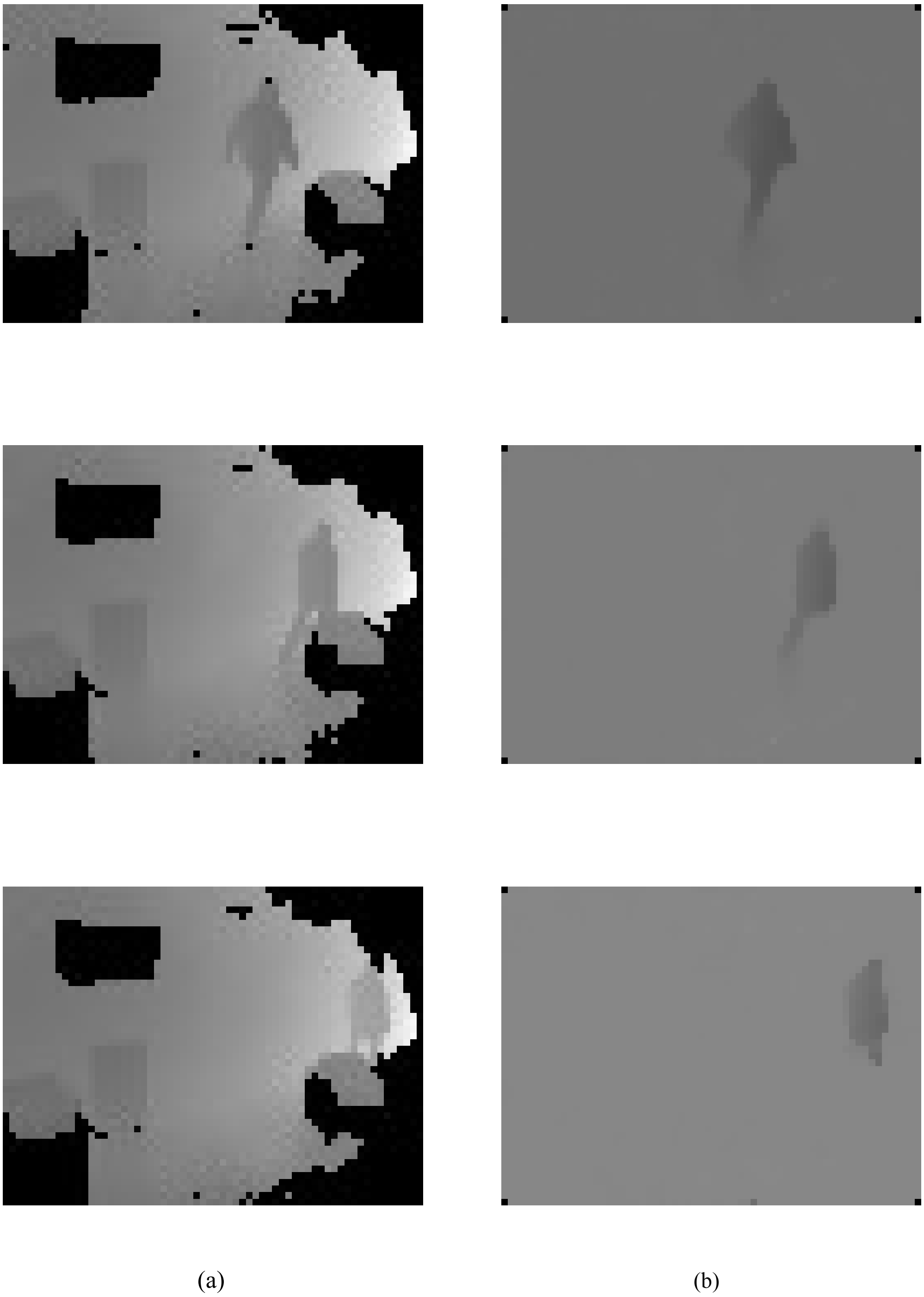


Figure 5.5 (a) Valid depth images after invalid measurement removal. (b) Human detection results from valid depth images.

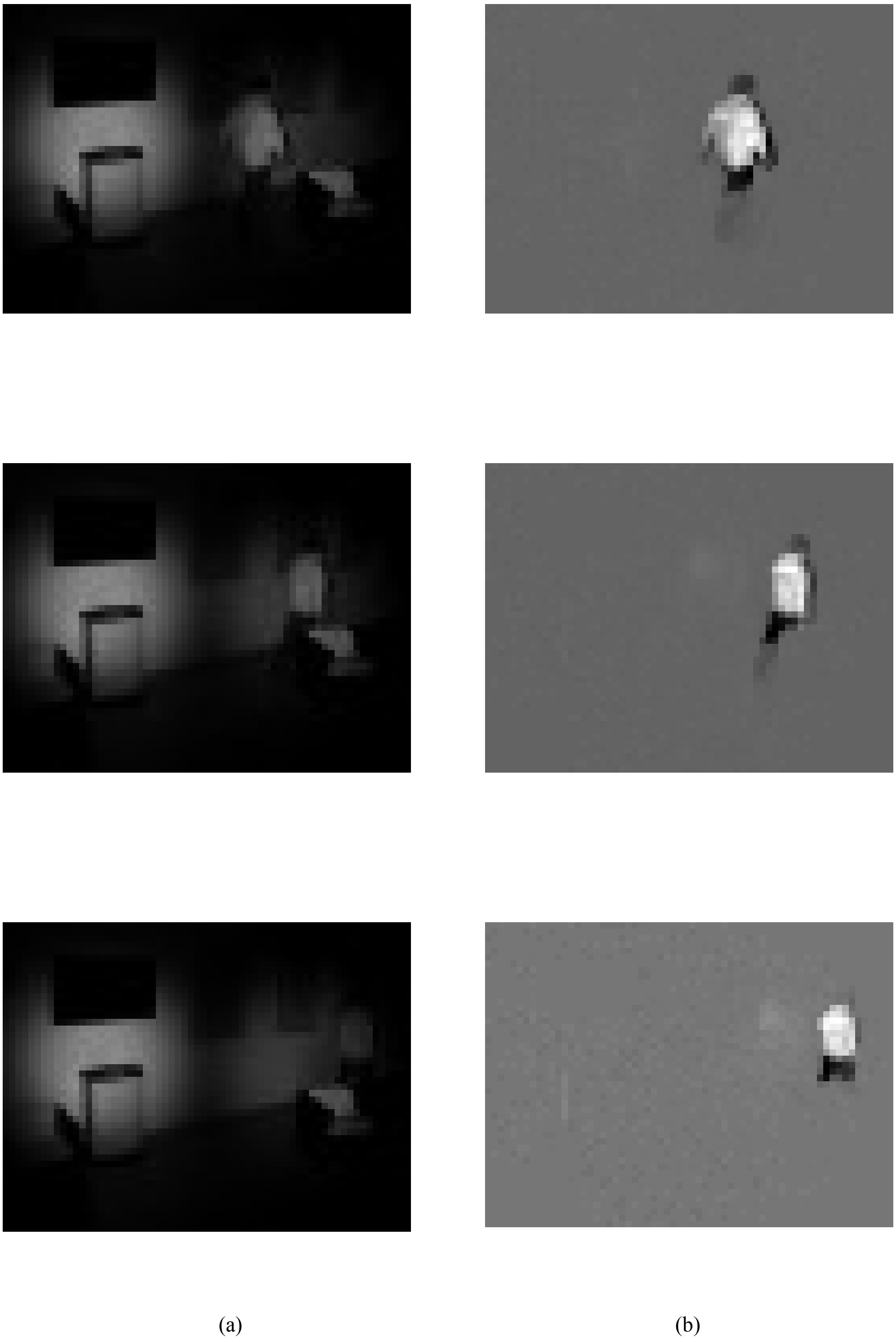


Figure 5.6 (a) Modulation amplitude. (b) Human detection results from modulation amplitude.

Sequence 2

For the application of indoor surveillance, it is quite common to have a scene room in a mess under bad illumination. In the image sequence the scene consists of an empty laboratory, which has two sets of fluorescent lights on the ceiling. All the lights are switched when outdoor lighting conditions are completely dark, and therefore the illumination could be caliginous and changed gradually. A person walks in and takes an object, finally walks out of the field of view of the MultiCam.

The random three frames from this sequence are presented in Figure 5.7 (a). As we mentioned before, we first down-sample the color images from pixels to pixels using simple down-scale technique. And the sequential 30 frames are trained for these multi-modal images. The balance parameter is set as $\lambda = 1$. The original color image and human detection results are shown in Figure 5.7 (a) and (b), respectively. The human is clearly detected although the environmental illumination is worse than the previous sequence, despite white line or part projected on his body. This is because that our model learns background template and human motion trajectories captured from the sequential 30 frames, and therefore, different color clusters which allows a quite fair discrimination among colors. However, as it can be seen from Figure 5.8 (a), the original depth map delivered from the MultiCam is quite highly noisy in the right part of the image due to invalid measurement in modulation amplitude map. The tradeoff parameter is fixed as $\lambda = 1$ and the results are presented in Figure 5.8 (b). In this case, however, we can observe that the human also is perfectly detected despite of high noise. In the meanwhile the detection results of the modulation amplitude map present a quite promising performance as shown in Figure 5.9 (b). Here we also set $\lambda = 1$.

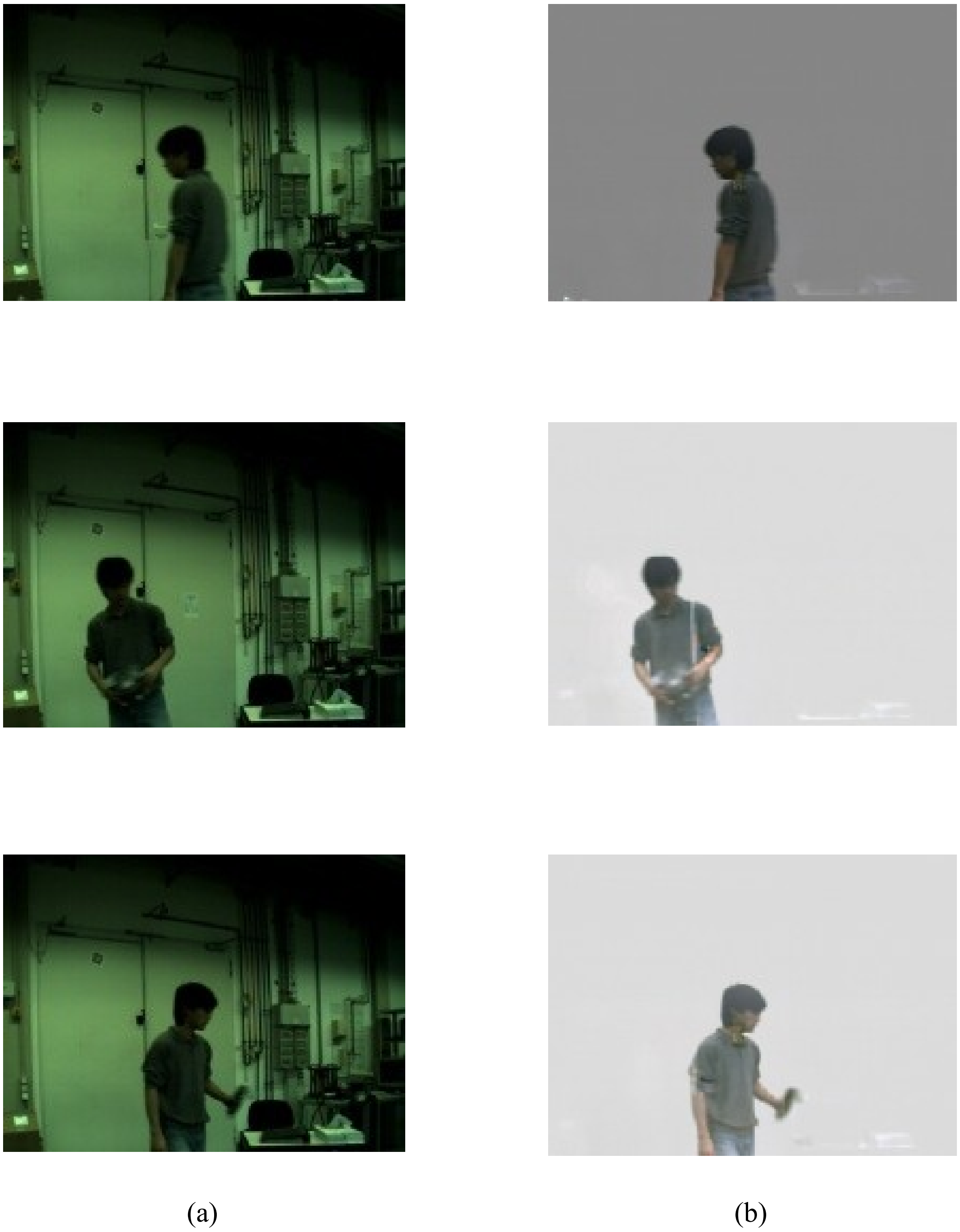
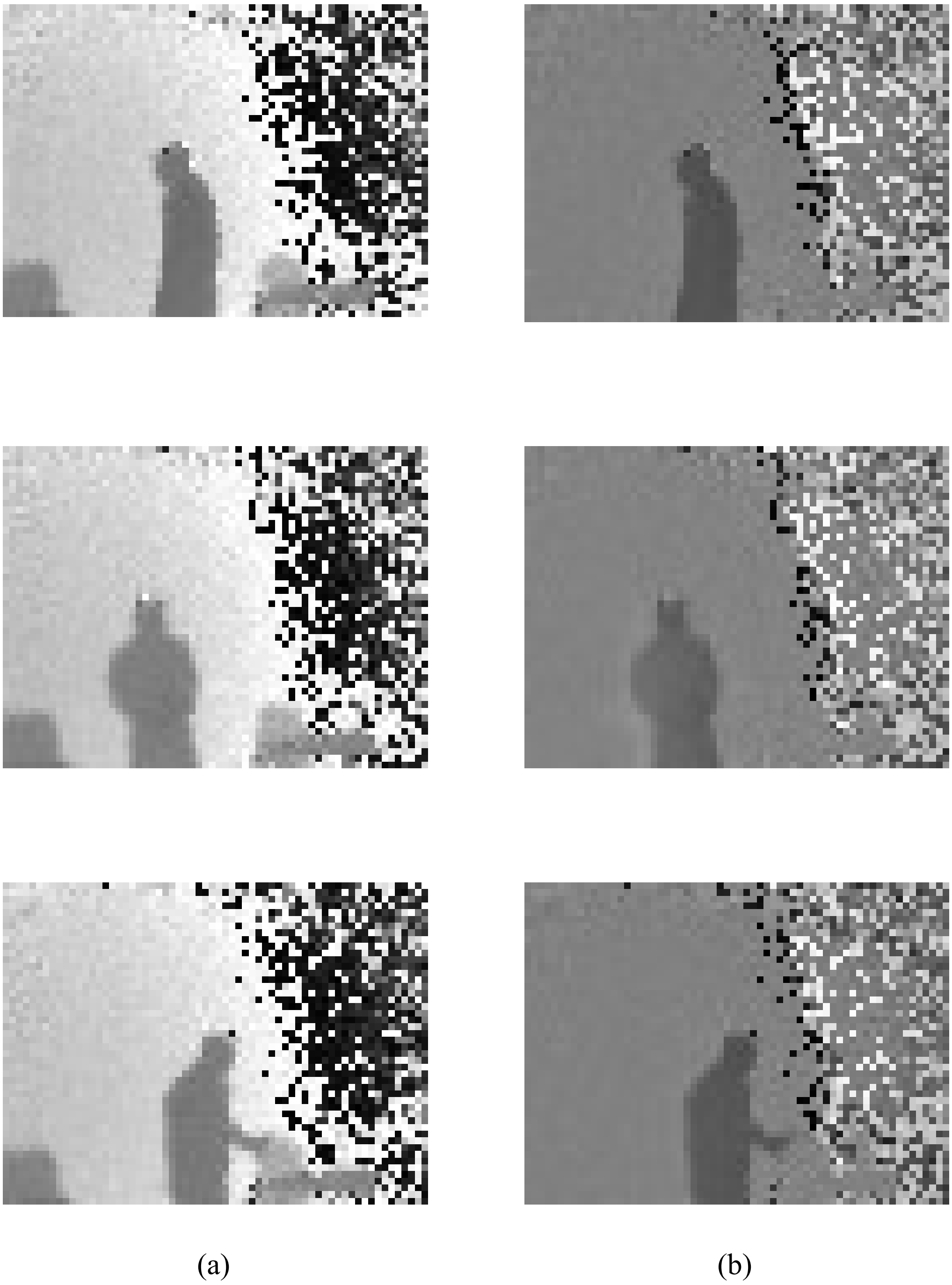


Figure 5.7 (a) Color image frames. (b) Human detection results.



(a) (b)

Figure 5.8 (a) Depth image frames. (b) Human detection results.

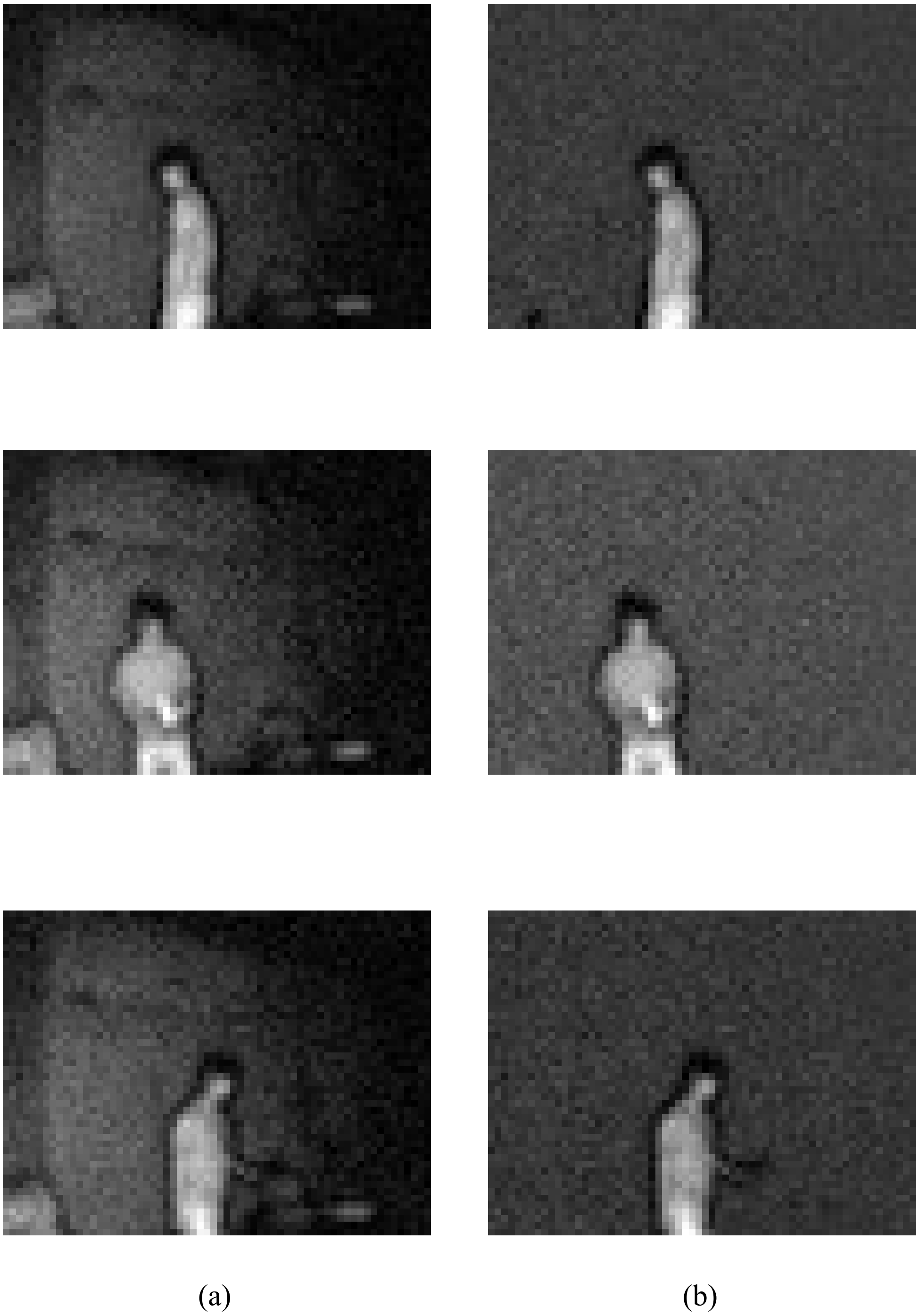


Figure 5.9 (a) Modulation amplitude images. (b) Human detection results.

5.5 Chapter summary

To take the advantages of multiple sensor of different modal, a 2D/3D vision system has been proposed by the researchers in recent years. As proposal of 2D/3D vision system, the MultiCam was developed recently as a novel combinational use of multi-cameras which incorporates a CMOS chip and a PMD ToF sensor in a monocular optical setup. And afterwards the setup of the MultiCam is described in detail.

We present a novel method towards human detection using sparse feature pursuit for multi-modal images simultaneously provided by a new monocular hybrid 2D/3D imaging system, which was developed based on the ToF principle in recent few years. The procedure of human detection in multi-modal images is implemented based on the following stages. We first make a conception that images with almost stationary background and dynamic foreground can be considered as the samples of signals that change slowly in time with the sparse feature with arbitrary shape caused by dynamic foreground. With the sparse prior, we cast a frame of video as a composition of two distinct layers: background and sparse foreground. And furthermore we recovery the two components based on the recent work on recovery of corrupted sparse low-rank matrix. Finally we reconstruct the foreground which contains human map from the sparse matrix. The experiments on the real image data on the one hand demonstrate the effectiveness of our proposed method, and on the other hand pave a promising way for the combination of multimodal images to yields a greatly improved detection result compared to either type of image data alone.

6 Summary and discussion

6.1 Summary

The broad objective of the dissertation is to investigates and explore the applications of the compressive sensing in multi-modal image processing and analysis. Three different kinds of applications are developed and presented. The first one presented an effective image fusion scheme based on the CS theory. The second one exploits the basis of CS to implement reconstruction of super-resolution of range/depth image. The last one aims to explore a potential approach based on CS to sense change in multi-modal images (i.e., 2D/3D images). The motivation, objective and the main contribution of the work to the related field are summarized in the first chapter. The conclusions of the dissertation are drawn as the following points:

- The combinational use of 2D color standard camera and 3D ToF camera provides a new perspective to sense the 3D real world due to it provides not only 2D color image but also 3D depth information.

- Along with a rapid increase in applications of difference, a rich theory of sparse and compressible signal reconstruction has recently been developed under the names of compressive sensing and sparse approximation. More recently, an offshoot of compressive sensing has become a focus of research on other low-dimensional signal structure such as matrix of low-rank and sparsity. The revolutionary research has inspired and initiated intriguing new research directions, and been contributing in related areas including image processing and signal processing. It provides an alternative to Nyquist-Shannon sampling theorem when the signal under acquisition is known to be sparse or compressible.

- The sparsity of signal or image is a ubiquitous property that plays a significant role in signal or image analysis. It has been shown in practice that various images of interest may be (approximately) represented sparsely to some extent and the sparse modeling is quite beneficial, or even crucial to solutions.

- Super-resolution is a software solution with the use of image processing algorithm which presumably is relatively inexpensive to implement in any situation where high-quality optical imaging system cannot be incorporated or too expensive to utilize.

- The super-resolution reconstruction is an ill-posed inverse problem. The problem model of super-resolution and the underlying solution can be treated from the perspective of compressive sensing strategy.

- The choice of sparsifying basis is very crucial in super-resolution reconstruction. And it is considered based on the property of image under some domain. The sensing matrix is extremely sensitive to the sparsifying basis according to the RIP property. In addition, the reconstruction algorithms also play a significant role in the result.

- As a natural extension of the standard sparsity concept in compressive sensing, the sparse feature can be considered as a cluster collected from the non-zero values in some practical applications.

- Sparse and low-rank matrix decomposition plays a prominent role in signal and image processing.

6.2 **Discussion and outlook**

Based on the current work presented in the thesis, a few topics for further discussion and study are outlined in which the following aspects are mainly concerned:

- MultiCam

The range measurement in the MultiCam which provides 2D/3D multimodal image is restricted. Thus, while the objects over the maximum distance can be observed in 2D image of the MultiCam, they have no any reliable depth information in the 3D image. Another one of the limitations is its poor performance in the situation of outdoor due to the ToF depth data are affected by the sun light to some extent. The current ToF sensors have no advantage in price in comparison with the conventional CCD and

CMOS sensors. However, with the development of ToF sensors technology, it is most possible to improve the performance of MultiCam by employing the new version of ToF sensor with higher resolution.

- Data fusion

The 2D color image and 3D depth image used in the thesis are processed via the proposed modeling. The obtained results represent only the 3D position change of the object. Fusion of the results from different modal can offer the full shape of moving object.

- Sparsifying basis

Although the depth map is sparsely represented under the wavelet basis, we suppose that there might be better-suited wavelet basis for the sparsifying basis. The sensing matrix is sensitive to the choice of the sparsifying basis according to the restrict isometry property, which means there seems to exist better sparsifying basis such as complex wavelets for sparse representation. In the meanwhile, the random point down-sample mask can be used to improve the effectiveness of the proposed method. In addition, the reconstruction algorithms also play an important role for the result. The other recover algorithms would be interesting to explore.

- Matrix recovery algorithms

Algorithms for the recovery of large-scale sparse and low-rank matrix influence the accuracy and efficiency of reconstruction result. It is believable that there might be better-suited recovery algorithms for the applications.

Bibliography

[1] E. J. Candes and M. B. Wakin, “An introduction to compressive sampling,” IEEE Signal Process. Mag., vol. 25, no. 2, pp.21–30, Mar. 2008.

[2] E. J. Candes and M. B. Wakin, “People hearing without listening: an introduction to compressive sampling,” IEEE Signal Processing Magazine. vol. 25, no. 2, pp.21-30, 2008.

[3] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by Basis Pursuit,” SIAM SCI. Comput., vol. 20, no. 1, pp.33–61, Aug. 1998.

[4] M. Lindner and A. Kolb, “Lateral and depth calibration of PMD distance sensors,” Advances in Visual Computing, volume 2, pp.524–533, Springer, 2006.

[5] Prasad, T., Klaus Hartmann, Wolfgang Weihs, Seyed Eghbal Ghobadi, and Arnd Sluiter. "First steps in enhancing 3D vision technique using 2D/3D sensors." In Computer Vision Winter Workshop, pp. 82-86. 2006.

[6] PMD Tech, 3D Video Sensor Array with Active SBI, 2009, website: http://www.pmdtec.com/

[7] M. Lindner and A. Kolb, “Calibration of the intensity-related distance error of the PMD TOF-camera,” SPIE: Intelligent Robots and Computer Vision XXV, volume 6764, pp. 6764–35, 2007.

[8] H. Nyquist, “Certain topics in telegraph transmission theory,” Transactions of the A.I.E.E., pp.617–644, 1928.

[9] C. E. Shannon, “Communication in the presence of noise,” Proc. Institute of Radio Engineers 37(1), pp.10–21, 1949.

[10] E. J. Candes, J. Romberg, and T. Tao, “Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information,” IEEE Trans. on Info. Theory, vol. 52, no. 2, pp.489–509, 2006.

[11] D. L. Donoho, “Compressed sensing,” IEEE Transactions on Information Theory, vol. 52, no. 4, pp.1289–1306, Apr. 2006.

[12] E. J. Candes and T. Tao, “Decoding by linear programming,” IEEE Trans. Inform. Theory 15(12), pp.4203–4215, 2005.

[13] E. J. Candes and T. Tao, “Near optimal signal recovery from random projections: universal encoding strategies?” IEEE Transactions on Information Theory, vol. 52, no. 12, pp.5406–5425, Dec. 2006.

[14] E. Candès, M. Rudelson, T. Tao, and R. Vershynin, “Error correction via linear programming,” Found. Comput. Math., pp. 295–308, 2005.

[15] S. Chen, D. Donoho, and M. Saunders, “Atomic decomposition by basis pursuit,” SIAM J. Sci. Comput., vol. 20, no. 1, pp.33–61, 1998.

[16] D. Needell and R. Vershynin, “Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit,” 2007, preprint.

[17] D. Needell and J. A. Tropp, “CoSaMP: Iterative signal recovery from incomplete and inaccurate samples,” Appl. Comput. Harmon. Anal.,vol. 26, no. 3, pp.301–321, 2008.

[18] J. A. Tropp and A. C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” IEEE Trans. Inf. Theory, vol. 53, no. 12, pp.4655–4666, Dec. 2007.

[19] E. Cand`es, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis,” Journal of the ACM , 58, 2011.

[20] S. Boyd and L. Vandenberghe, “Convex Optimization,” Cambridge University Press, 2004.

[21] M. Fazel, “Matrix Rank Minimization with Applications,” PhD thesis, Stanford University, 2002.

[22] Wen Cao, Bicheng Li and Yong Zhang, “A remote sensing image fusion method based on PCA transform and wavelet packet transform,” Proceedings of the 2003 International Conference on Neural Networks and Signal Processing, Dec. 2003.

[23] Sveinsson, Johannes R., Magnus Orn Ulfarsson, and Jon Atli Benediktsson, “Cluster-based feature extraction and data fusion in the wavelet domain,” IEEE International Geoscience and Remote Sensing Symposium, vol. 2, pp. 867-869, 2001.

[24] S. Mallat, “A Wavelet Tour of Signal Processing,” Academic Press, Second Edition, 1998.

[25] Garzelli, A, “Possibilities and Limitations of the Use of Wavelets in Image Fusion,” Pro. IEEE International Geoscience and Remote Sensing Symposium, 2002.

[26] Wen Doua, Yunhao Chen, “An Improved Image Fusion Method with High Spectral Fidelity,” The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B7. Beijing 2008.

[27] Lau Wai Leung, Bruce King and Vijay Vohora, “Comparison of Image Fusion Techniques using Entropy and INI,” Pro. 22nd Asian Conference on Remote Sensing, Nov. 2001.

[28] Ghantous, Milad, Soumik Ghosh, and Magdy Bayoumi. "A gradient-based hybrid image fusion scheme using object extraction." Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on. IEEE, 2008.

[29] Z. Lin, M. Chen, L. Wu, and Y. Ma, “The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices,” UIUC Technical Report UILU-ENG-09-2215, November 2010.

[30] Donoho, David L., et al., “Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit.” Information Theory, IEEE Transactions on 58.2 (2012): pp.1094-1121, 2012.

[31] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, “Fast Convex Optimization Algorithms for Exact Recovery of a Corrupted Low-Rank Matrix,” UIUC Technical Report UILU-ENG-09-2214, August 2009.

[32] Cai, Jian-Feng, Emmanuel J. Candès, and Zuowei Shen, “A singular value thresholding algorithm for matrix completion,” SIAM Journal on Optimization 20, no. 4 (2010): pp.1956-1982, 2010.

[33] Lange R. and Seitz P., “Solid-state Time-Of-Flight Range Camera,” IEEE Journal of Quantum Electronics, Vol.37, pp.390-397, 2001.

[34] PMD Tech, 3D Video Sensor Array with Active SBI, http://www.pmdtec.com/, 2009.

[35] Moeler, T., Kraft, H., Frey, J., Albrecht, M. and Lange, R., “Robust 3D Measurement with PMD Sensors,” Range Imaging Day, Zürich, 2005.

[36] Lin, Zhouchen, Arvind Ganesh, John Wright, Leqin Wu, Minming Chen, and Yi Ma, “Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix,” Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP) 61, 2009.

[37] S. E. Ghobadi, O. E. Loepprich, K. Hartmann, and O. Loffeld, “Hand segmentation using 2d/3d images,” IVCNZ 2007 Conference, Hamilton, New Zealand, 2007.

[38] S. E. Ghobadi, O. E. Loepprich, O. Lottner, K. Hartmann, O. Loffeld, and W. Weihs, “Improved object segmentation based on 2d/3d images,” The Fifth IASTED International Conference on Signal Processing, Pattern Recognition, and Applications (SPPRA 2008), pages 42–47. ACTA Press, 2008.

[39] J. Leens, S. Pi´erard, O. Barnich, M. V. Droogenbroeck, and J.-M.Wagner, “Combining color, depth, and motion for video segmentation,” in ICVS, pp. 104–113, 2009.

[40] C. Stauffer and W. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” IEEE Computer SocietyConference on Computer Vision and Pattern Recognition, vol. 2, pp.252, 1999.

[41] L. Bianchi, P. Dondi, R. Gatti, L. Lombardi, and P. Lombardi, “Evaluation of a foreground segmentation algorithm for 3d camera sensors,” ICIAP, ser. Lecture Notes in Computer Science, vol. 5716. Springer, pp. 797–806, 2009.

[42] Michael Harville, Gaile Gordon, John Woodfill, “Foreground Segmentation Using Adaptive Mixture Models in Color and Depth,” event, pp.3, IEEE

Workshop on Detection and Recognition of Events in Video (EVENT'01), 2001.

[43] J. Diebel and S. Thrun, “An application of markov random fields to range sensing,” Proceedings of Conference on Neural Information Processing Systems (NIPS), Cambridge, MA, MIT Press, 2005.

[44] J. Kopf, M. Cohen, D. Lischinski, and M. Uyttendaele, “Joint bilateral upsampling,” ACM TOG, 26(3), 2007.

[45] Q. Yang, R. Yang, J. Davis, and D. Nister, “Spatial-depth super resolution for range images,” In IEEE Computer Vision and Pattern Recognition. IEEE Computer Society, 2007.

[46] Tomasi, C. and Manduchi, R., “Bilateral filtering for gray and color images,” Proc. IEEE ICCV, pp.839 –846 , 1998.

[47] H. S. Hou and H. C. Andrews, “Cubic splines for image interpolation and digital filtering,” IEEE Trans. on SP, 26(6):5 pp.08–517, 1978.

[48] P. Thevenaz, T. Blu, and M. Unser, “Image Interpolation and Resampling,” Handbook of Medical Imaging, Processing and Analysis, Academic Press, San Diego, USA, 2000.

[49] Park, J., Kim, H., Tai, Y.W., Brown, M., and Kweon, I., “High quality depth map upsampling for 3D-ToF cameras,” ICCV. 2011.

[50] R. Crabb, C. Tracey, A. Puranik, and J. Davis, “Real-time foreground segmentation via range and color imagin,”. Proc. of CVPR Workshop on Time-of-flight Computer Vision, 2008.

[51] Irani, M., Peleg, S., “Improving resolution by image registration,” CVGIP: Graph. Models Image Process. 1991.

[52] Schuon, S., Theobalt, C., Davis, J., and Thrun, S., “Lidarboost: Depth superresolution for tof 3D shape scanning,” CVPR. 2009.

[53] G. Malathi and V. Shanthi, “Wavelet Image Fusion Approach for Classification of Ultrasound Placenta Complicated by Gestational Diabetes Mellitus, Pathophysiology and Complications of Diabetes Mellitus,” Prof. Oluwafemi Oguntibeju (Ed.), ISBN: 978-953-51-0833-7, InTech, DOI: 10.5772/53530. 2012.

[54] Hahne, U., and Alexa, M., “Exposure Fusion for Time-Of-Flight Imaging,” Pacific Graphics, 2011.

[55] Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R.A., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A.J., and Fitzgibbon, A., “Kinectfusion: Real-time 3D reconstruction and interaction using a moving depth camera,” UIST. 2011.

[56] Cui, Y., Schuon, S., Derek, C., Thrun, S., and Theobalt, C., “3D shape scanning with a time-of-ight camera,” CVPR. 2010.

[57] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," Image processing, IEEE Transactions on, vol. 13, no. 10, pp. 1327–1344, Oct. 2004.

[58] E. Candès, M. Rudelson, T. Tao, and R. Vershynin, "Error correction via linear programming," Found. Comput. Math., pp. 295–308, 2005.

[59] D. P. Bertsekas, "Nonlinear Programming," Athena Scientific, 2004.

[60] Luan, Xuming, "Experimental investigation of photonic mixer device and development of TOF 3D ranging Ssystems based on PMD technology," 2006.

[61] Haindl, M.; Zid, P., "Multimodal Range Image Segmentation, Vision Systems: Segmentation and Pattern Recognition," 2007.

[62] Song Zhang, Peisen Huang, "High-resolution, real-time 3-D shape measurement," Optical Engineering: 123601. 2006.

[63] P. Palojärvi, "Integrated Electronic and optoelectronic circuits and devices for pulsed time-of-flight laser rangefinding," University of Oulu, Finland, 2003.

[64] Ari Kilpelä, "Pulsed time-of-flight laser ranfe finder techniques for fast, high precision measurement applications," University of Oulu, Finland, 2004.

[65] J. Takeno and U. Rembold, "Stereo vision system for autonomous mobile robot," Intelligence Autonomous Systems. IAS-4: Proceedings of the Internattional Conference, Karlsruhe, Germany, March pp.27-30, 1995.

[66] Klimentjew, Denis, Norman Hendrich, and Jianwei Zhan, "Multi sensor fusion of camera and 3D laser range finder for object recognition." Multisensor Fusion and Integration for Intelligent Systems (MFI), 2010 IEEE Conference on. IEEE, 2010.

[67] Palojärvi, Pasi, "Integrated electronic and optoelectronic circuits and devices for pulsed time-of-flight laser rangefinding," Oulun yliopisto, 2003.

[68] Aliakbarpour, Hadi, P. Nuez, Jose Prado, Kamrad Khoshhal, and Jorge Dias, "An efficient algorithm for extrinsic calibration between a 3d laser range finder and a stereo camera for surveillance," Advanced Robotics, 2009. ICAR 2009. International Conference on, pp. 1-6, IEEE, 2009.

[69] Joung, Ji Hoon, et al., "3D environment reconstruction using modified color ICP algorithm by fusion of a camera and a 3D laser range finder," Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on. IEEE, 2009.

[70] Monteiro, Gonçalo, et al., "Tracking and classification of dynamic obstacles using laser range finder and vision," Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2006.

[71] Van den Bergh, Michael, and Luc Van Gool., "Combining RGB and ToF cameras for real-time 3D hand gesture interaction," Applications of Computer Vision (WACV), 2011 IEEE Workshop on. IEEE, 2011.

[72] Ghobadi, Seyed Eghbal, Omar Edmond Loepprich, Oliver Lottner, W. Weihs, Klaus Hartmann, and O. Loffeld, “Analysis of the Personnel Safety in a Man-Machine Cooperation Using 2D/3D Images,” Proceedings of the EURON/IARP International Workshop on Robotics for Risky Interventions and Surveillance of the Environment. 2008.

[73] Netramai, C., O. Melnychuk, C. Joochim, and H. Roth. "Combining pmd and stereo camera for motion estimation of a mobile robot." Center for Sensor Systems (ZESS). Paul-Bonatz-Strasse 9, no. 11 (2008): 57076.

[74] Hahne, Uwe, and Marc Alexa, “Depth imaging by combining time-of-flight and on-demand stereo,” Dynamic 3D Imaging, pp. 70-83. Springer Berlin Heidelberg, 2009.

[75] Structured-light 3D scanner, http://en.wikipedia.org/wiki/Structured-light_3D_scanner.

[76] Furht, Borko, ed. “Encyclopedia of multimedia,” Springer-Verlag New York Incorporated, 2008.

[77] Rick S. Blum, and Zheng Liu, “Multi-Sensor Image Fusion and Its Applications,” Published by: CRC Press. 2005.

[78] D.L.Donoho, “Compressed sensing,” IEEE Transactions on Information Theory, 2006, 52(4): pp.1289-1306, 2006.

[79] E.Candes, “Compressive sampling,” Proceedings of International Congress of Mathematicians. Zurich, Switzerland: European Mathematical Society Publishing House, pp.1433-1452, 2006.

[80] T. Wan, N. Canagarajah, and A. Achim, “Compressive image fusion,” Proceeding of 15th IEEE International Conference on Image Processing. San Diego, California, U.S.A: IEEE, pp.1308 – 1311, 2008.

[81] X.Luo, J.Zhang, J.Yang, and Q.Dai, “Image fusion in compressive sensing,” Proceeding of 16th IEEE International Conference on Image Processing.Cairo, Egypt: IEEE, pp.2205-2208, 2009.

[82] Y.Tsaig and D.L.Donoho, “Extensions of compressed sensing”, Signal Processing, 86(3): pp.549-571, 2006.

[83] Duarte, Marco, Michael Wakin, Shriram Sarvotham, Dror Baron, and Richard G. Baraniuk, “Distributed compressed sensing of jointly sparse signals,” 2005.

[84] S. Mallet and Z. Zhang, “Matching pursuit in a time-frequency dictionary,” IEEE Transactions on Signal Processing, pp.3397–3415, 1993.

[85] Tibshirani, Robert, “Regression shrinkage and selection via the lasso,” Journal of the Royal Statistical Society. Series B (Methodological) (1996): pp.267-288, 1996.

[86] Chen, Shaobing, and David Donoho, “Basis pursuit,” Signals, Systems and Computers, 1994. 1994 Conference Record of the Twenty-Eighth Asilomar Conference on. Vol. 1. IEEE, 1994.

[87] Candes, Emmanuel J., and Terence Tao, “Decoding by linear programming,” Information Theory, IEEE Transactions on 51.12 (2005): pp.4203-4215, 2005.

[88] Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani, “Least angle regression,” The Annals of statistics 32, no. 2 (2004): pp.407-499, 2004.

[89] Figueiredo, Mário AT, Robert D. Nowak, and Stephen J. Wright, “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems,” Selected Topics in Signal Processing, IEEE Journal of 1.4 (2007): pp.586-597, 2007.

[90] Daubechies, Ingrid, Michel Defrise, and Christine De Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint,” Communications on pure and applied mathematics 57.11 (2004): pp.1413-1457, 2004.

[91] W. Yin, S. Osher, J. Darbon, and D. Goldfarb, “Bregman Iterative Algorithms for Compressed Sensing and Related Problems,” SIAM Journal on Imaging Sciences, 1(1): pp.143-168, 2008.

[92] Hale, Elaine T., Wotao Yin, and Yin Zhang, “A fixed-point continuation method for l1-regularized minimization with applications to compressed sensing,” CAAM TR07-07, Rice University, 2007.

[93] Wen, Zaiwen, et al., “A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation,” SIAM Journal on Scientific Computing 32.4 (2010): pp.1832-1857. 2010.

[94] Gordon, G., Trevor Darrell, Michael Harville, and John Woodfill, “Background estimation and removal based on range and color,” Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., vol. 2. IEEE, 1999.

[95] Plagemann, C., Ganapathi, V., Koller, D., Thrun, S., “Real-time identify ication and localization of body parts from depth images,” Proc. of ICRA'2010, pp.3108-3113, 2010.

[96] Rajagopalan, A. N., Arnav Bhavsar, Frank Wallhoff, and Gerhard Rigoll, “Resolution enhancement of pmd range maps,” Pattern Recognition, pp. 304-313. Springer Berlin Heidelberg, 2008.

[97] Schuon, S., Theobalt, C., Davis, J. and Thrun, S., “High-quality scanning using time-of-flight depth superresolution,” Computer Vision and Pattern Recognition Workshops, 2008.CVPRW ’08. IEEE Computer Society Conference on, pp. 1–7. 2008.

[98] Gonzalez-Banos, Hector, and James Davis. "Computing depth under ambient illumination using multi-shuttered light." CVPR 2004.

[99] B. Jutzi and U. Stilla, “Precise range estimation on known surfaces by analysis of full-waveform laser,” Proceedings of Phtogrammetric Computer Vision PCV, 2006.

[100] M. Laurenzis, F. Christnacher, and D. Monnin, “Long-range three-dimensional active maging with superresolution depth mapping,” Opt. Lett., 32(21): pp.3146–3148, 2007.

[101] Park, Sung Cheol, Min Kyu Park, and Moon Gi Kang, “Super-resolution image reconstruction: a technical overview,” Signal Processing Magazine, IEEE 20 no.3 (2003): pp.21-36, 2003.

[102] Chui, Charles K, “An introduction to wavelets,” Vol. 1. Access Online via Elsevier, 1992.

[103] P. J. Burt and E. H.Adelson, “The Laplacian pyramid as a compact image code,” Proc.IEEE Int. Conf.Commun., 31(4): pp.532 – 540, 1983.

[104] S. Mallat, “A theory for multiresolution signal decomposition: The wavelet representation,” IEEE Trans.Patt.Anal. Mach. Intell., 11(7): pp.674 – 693, 1989.

[105] S. Mallat, “Multiresolution approximations and wavelet orthonormal bases of L2,” Trans.Amer. Math. Soc., 315: pp.69 – 87, 1989.

[106] Daubechies, Ingrid, “Ten lectures on wavelets. Society for Industrial and Applied Mathematics, 1992.” Rulgers University and AT&T Bell Laboratories 2009.

[107] Chen, Chia-Ping, Yu-Ting Chen, Ping-Han Lee, Yu-Pao Tsai, and Shawmin Lei, “Real-time hand tracking on depth images,” Visual Communications and Image Processing (VCIP), 2011 IEEE, pp. 1-4, IEEE, 2011.

[108] Schoenberg, Jonathan R., Aaron Nathan, and Mark Campbell, “Segmentation of dense range information in complex urban scenes,” Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on. IEEE, 2010.

[109] Candes, Emmanuel J., Yonina C. Eldar, Deanna Needell, and Paige Randall, “Compressed sensing with coherent and redundant dictionaries,” Applied and Computational Harmonic Analysis 31, no. 1: pp. 59-73, 2011.

[110] Candes, Emmanuel, and Justin Romberg, “Sparsity and incoherence in compressive sampling.,” Inverse problems 23, no. 3: pp.969, 2007.

[111] Rocchini, C. M. P. P. C., Paulo Cignoni, Claudio Montani, Paolo Pingi, and Roberto Scopigno, “A low cost 3D scanner based on structured light,” In Computer Graphics Forum, vol. 20, no. 3, pp. 299-308. Blackwell Publishers Ltd, 2001.

[112] Geng, Jason, “Structured-light 3D surface imaging: a tutorial,” Advances in Optics and Photonics 3, no. 2: pp.128-160, 2011.

[113] Bos, Philip J. "Stereoscopic imaging system with passive viewing apparatus." U.S. Patent 4,719,507, issued January 12, 1988.

[114] Lange, Robert, and Peter Seitz, “Solid-state time-of-flight range camera,” Quantum Electronics, IEEE Journal of 37, no. 3: pp.390-397, 2001.

Relevant publications

A.1 "Multi image fusion based on compressive sensing", International Conference on Audio, Language and Image Processing, IEEE, 2010.

Multi image fusion based on compressive sensing

Juanjuan Han, Otmar Loffeld, Klaus Hartmann, and Robert Wang
Center for Sensorsystems (ZESS), University of Siegen
{Han, Loffeld, Hartmann, wang}@zess.uni-siegen.de

Abstract

Compressive sensing provides a novel framework to acquire and to reconstruct a signal or digital image from sparse measurements acquired at sub-Nyquist/Shannon sampling rate. In this paper, we present an effective image fusion scheme based on a Discrete Cosine Transform (DCT) sampling model for compressive sensing imaging. A sparse sampling model according to the DCT-based spectral energy distribution is proposed. The compressive measurements of multiple input images obtained with the proposed sampling model are fused to a composite measurement by combining their wavelet approximation coefficients and their detail coefficients separately. The combination is done by applying a weighting operation for every sampling location according to the statistical distribution. Furthermore, the fused image is reconstructed from the composite measurement by solving a problem of total variation minimization. Finally, we validate the effectiveness of the algorithm using multiple images.

1. Introduction

With the rapid development of sensor systems, the information science focuses mainly on how the information about the real world is extracted from the sensor data. In many cases, a single sensor is not sufficient to provide a complete and fully informative perception of the real world. Therefore, multi-sensor fusion has attracted a great deal of attention in the past years. Image fusion is a branch of multi-sensor fusion and refers to a process of combining relevant information from two or more images into a fused image that possesses more information than any of the input images [1]. The current image fusion schemes can be classified roughly into pixel-based and region-based methods. For both of them all the samples of the images have to be acquired, which means that the storage burden and the processing challenges must be handled especially due to the growing sensor data volumes. Recently, an exciting new field, known as compressive sensing (CS), establishes mathematically that a relatively small number of non-adaptive, linear measurements can harvest all of the information necessary to faithfully reconstruct sparse or compressible signals [2]-[3]. The CS theory exploits the knowledge that the signal or image we are acquiring is sparse in some known transform domain, which means that the signal or image is compressive [2]-[3]. Then the compressive signal may be reconstructed accurately with sub-Nyquist/Shannon data sampling rate from a significantly smaller number of measurements than sampling the original signal at Nyquist/Shannon rate [2]-[3]. This is a clear and striking advantage compared with the conventional signal theory based on the Shannon theory. Therefore, the CS theory can lead to the reduction of sampling rates, storage volume, power consumption, and computational complexity in signal and image processing and related research fields.

Regarding image fusion in CS, one natural way is to fuse the images after being reconstructed from the random projections. However, in order to reduce the computational complexity and to save storage space, a better way is to directly combine the measurements in the compressive domain, and then to reconstruct the fused image from the fused measurements. There are several different methods which have been proposed in recent years, e.g., a simple maximum selection fusion rule [4] or a weighted average based on entropy metrics of the original measurements [5].

In image compression, due to its computational simplicity and the fact that the spectral coefficients are real numbers, the Discrete Cosine Transformation (DCT) rather than the Fast Fourier Transformation (FFT) is widely used to represent a signal sparsely. The advantage of dealing with real rather than complex numbers also simplifies the algorithmic implementation of compressive approaches conceptually. In this paper we propose therefore a DCT-based sampling model based on the sparsity of the image in the spectral domain. Sampling is performed on multiple input images using the proposed sampling model to obtain their linear measurements in the compressive domain. Inspired by the wavelet based fusion method, we propose a fusion scheme which combines the wavelet approximation coefficients and detail coefficients of the linear measurement series respectively via an energy distribution based weighting operation.

Finally, we reconstruct the fused image with total variance minimization algorithm [5].

This paper is organized as follows. In Section 2, we introduce the sparse digital image model and an overview of the CS background. The proposed fusion scheme based on CS is described in Section 3. Some experimental results and an analysis are provided in Section 4. Finally, Section 5 ends this paper with a conclusion.

2. Sparse model and compressive sensing background

2.1. Signal and digital image sparsity

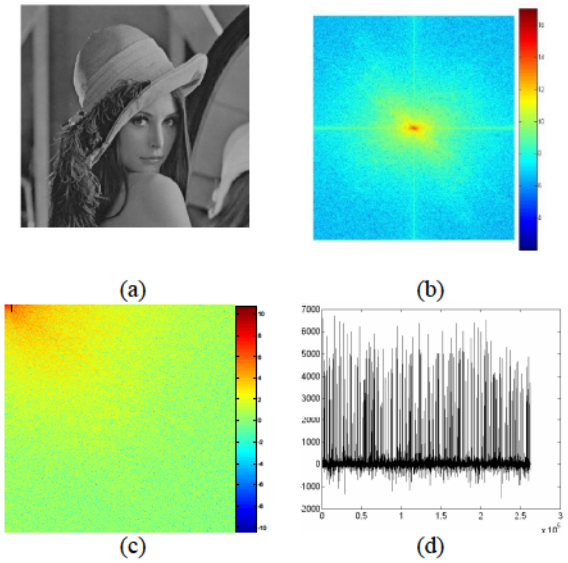
Let f be a vector signal $f \in R^N$. Then it can be represented with an orthogonal basis $\{\psi_i\}_{i=1}^N$ as: $f = \sum_{i=1}^{N} x_i \psi_i$, where $\{x_i\}$ denotes the coefficient sequence of f using the basis $\{\psi_i\}_{i=1}^N$. This representation can be formulated easily in matrix vector notation as:

$$f = \Psi x \tag{1}$$

where Ψ is the $n \times n$ matrix with ψ_i as columns. x is defined as the coordinate vector of f relative to Ψ, i.e., $x = (x_1, x_2, \cdots, x_N)^T$.

The signal f has a sparse representation if many coefficients in the vector x are small and therefore can be neglected without seriously degrading the signal. If K is the number of coefficients which are considerably larger than the rest, neglecting the $N-K$ small coefficients will hence not lead to a perceptual loss of the original signal [3]. Some examples of natural signals that have the sparse property have been identified such as digital images, radar pulse returns of a sparse scene, video sequences, and so on. In each of these cases the relevant information in a sparse representation of a signal is encoded in both the indices of the significant coefficients and the values of these.

The image data can be mapped to a sparse vector via a sparsifying transform. Different types of images have sparse representations under different transforms. Real-world images are known to have a sparse representation in the FFT, DCT and wavelet transform domain. The digital image "Lena" and its frequency transforms are shown in Fig.1.



(a) (b)

(c) (d)

Fig.1 (a) Original image. (b) Its FFT on log-scale (zero frequency in the center of the image). (c) Its DCT on log-scale (zero frequency in the upper left corner). (d) Wavelet coefficients.

In Fig.1 (a), nearly all pixels values in the original image are non-zero. However, the image tends to concentrate its energy in the frequency domain, where most energy concentrates at low frequencies or at a few large coefficients. Fig.1 (b) shows the FFT of this original image on log-scale with shifting the zero-frequency component to the center of the image. The low-frequency components in an image are normally much larger in amplitude than the high-frequency components. The DCT relocates the compact energy in the upper left corner of the image [6]. Lesser energy or information is distributed over other areas, as shown in Fig.1 (c). The image is converted to a sparse vector in DCT domain. Most information of the original image is concentrated statistically in just a few large coefficients, while most of the high frequency coefficients are either zero or close to zero. Similarly, an image can be represented by just a few large coefficients in the wavelet transform domain, as shown in Fig.1 (d). Thus, it can be said that the image has the sparsity property with a few large coefficients carrying most information using some orthogonal basis.

Recent literature on Compressive Sensing states that a signal may be reconstructed accurately from a small set of measurements if it is sparse in some orthogonal basis. This provides the possibility to directly process the measurements of multi-images in compressive sensing and then to reconstruct the fused image from the measurement according to a recovery algorithm such as Gradient Projection for Sparse Reconstruction (GPSR) [15], Orthogonal Matching Pursuit (OMP) [14], L1-norm minimization, total variation minimization [9], and so forth.

2.2. Overview of Compressive Sensing

We begin by revisiting the problem of recovering the signal f from a set of M measurements. Mathematically speaking, this compressive measurement vector can be formulated as

$$y = \Phi f \qquad (2)$$

where $y \in R^M (K < M << N)$, $\Phi \in R^{M \times N}$ is a measurement matrix (or an observation matrix in terms of state space theory). Since $M << N$, the recovery of the signal vector f from the measurement vector y is a highly underdetermined problem in general. However, the CS theory reveals that a signal can be reconstructed from the M measurements if the following conditions hold:

(1) The signal f can be represented sparsely by an orthogonal basis Ψ , shown in Eq.1.

(2) The orthogonal basis Ψ and compressive measurement matrix Φ are incoherent.

Satisfying the two conditions above, the signal can be recovered by solving an L_1 -minimization problem [7]-[10]:

$$\hat{x} = \arg\min \| x \|_1 \text{ s.t. } \tilde{\Phi} x = y \qquad (3)$$

where $\tilde{\Phi}$ is defined as

$$\tilde{\Phi} = \Phi \Psi \qquad (4)$$

Therefore, an estimate $\hat{f}$ of the signal can be recovered from $\hat{x}$ by the following equation:

$$\hat{f} = \Psi \hat{x} \qquad (5)$$

3. Image fusion scheme

3.1. Sampling

For the reconstruction of the fused image, we first construct a viable sensing matrix $\tilde{\Phi}$ which must satisfy the Restricted Isometry Property (RIP) property [9]. There are two ways to achieve this: (1) directly construct the sensing matrix $\tilde{\Phi}$ to follow this property; and (2) reduce the problem to a known matrix Φ that satisfies the RIP property. Examples are the random Gaussian matrix [13][16], the uniform Spherical ensemble, Random Partial Fourier matrices [11], Toeplitz matrix[12], and so forth.

The random partial Fourier matrix is used to expand the applicability of compressive sensing to large scale data such as 2D images due to the special structure of the Fourier transform under the partial Fourier ensemble [11]. Inspired by this work, we propose a sampling model designed to account for the property of the DCT in the frequency domain as shown in Fig. 2. The DCT relocates the energy of a digital image in the frequency domain. Most of the energy of a digital image concentrates at low frequencies (upper left corner shown in Fig. 1(c)). Hence most information of an image can found in the measurements located at the upper left corner of the image in the DCT domain. The sampling model contains many radial lines extending from the upper left corner to the other side of an image, as shown in Fig. 2. The measurement matrix Φ then is constructed from the sampling pattern in the 2D discrete cosine plane created by using nearest neighbor techniques.

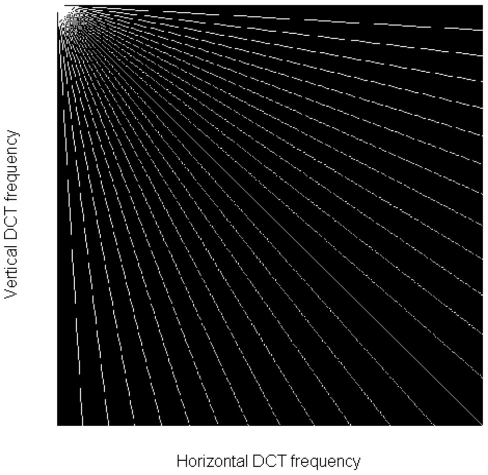


Fig. 2 DCT based sampling model (zero frequency in the upper left corner).

3.2. Fusion

For most of the conventional fusion approaches, the image fusion is performed on the level of the source images. With the emergence of CS theory, however, the fusion procedure can be implemented in the compressive domain. That is, to combine the individual linear measurement of multi-input images to a single composite measurement and to reconstruct the fused image from these composite measurement.

Consider a natural image with the size of $n \times n$. To simplify the complexity of the computation, the image data is usually arranged in a one dimensional column vector f of length $N (N = n \times n)$. We already know that the column vector f is sparse in the orthogonal 2D DCT basis Ψ according to Fig. 1(c). The measurement vector y (the DCT spectrum) is the projection of image onto the column vectors of the measurement matrix Φ . Mathematically speaking, the relationship can be expressed as $y = \Phi f = \Phi \Psi x = \tilde{\Phi} x$.

Suppose there are P images with the same size $n \times n$ need to be fused. After transforming them to 1D column vectors $f_p (p = 1, 2, ..., P)$ and supposing that all the images have similar spectral features, all the vectors will have compact representations in terms of the significant coefficients in the orthogonal basis Ψ. We collect the corresponding linear measurements $y_p (p = 1, 2, ..., P)$ with length $M (M < N)$ in one large augmented observation vector of size $M \times P$ (rather than $N \times P$). Thus the measurements are not the simple pixel values of the original images any more. The fusion among the original images can be considered naturally as the fusion among the linear measurements of $y_p (p = 1, 2, ..., P)$, which contain the important information reflecting the image texture.

Multi-Scale wavelet decomposition shows remarkable advantages in the representation of a signal. In this fusion scheme, we apply a single-level 1D Daubechies wavelet transform to decompose the linear measurement vectors into two components: approximation coefficients $A_p (p = 1, 2, ..., P)$ and detail coefficients $D_p (p = 1, 2, ..., P)$. The larger a coefficient is, the more information it carries. Therefore, a weighted mean is applied to incorporate the contributions of all inputs so that data elements with a high weight contribute more to the weighted mean than elements with a low weight. The fused approximation coefficient A and detail coefficient D can be formulated as:

$$A = \sum_{p=1}^{P} \alpha_p A_p \qquad (6)$$

$$D = \sum_{p=1}^{P} \beta_p D_p \qquad (7)$$

where α_p and β_p are the weighting factors, and defined as

$$\alpha_p = \frac{|A_p|}{\sum_{p=1}^{P} |A_p|} \qquad \beta_p = \frac{|D_p|}{\sum_{p=1}^{P} |D_p|} \qquad (8)$$

Consequently the fused linear measurement y can be obtained through the inverse discrete wavelet transform. This is a process by which components can be assembled back into the original signal without loss of information. Finally, the 1D column vector f of the fused image is reconstructed from the fused linear measurement y via the recovery algorithm total variation minimization, see [9].

4. Experimental results and performance evaluation

In this section, we perform two groups of comparisons for the performance evaluation to illustrate the effectiveness of the proposed approach. In the experiments all the input images have the sparsity property in the 2D discrete cosine transform domain. The fused images are reconstructed from $M = N / 2$ measurements. In this paper, we compare the proposed scheme with the maximum selection fusion rule in [4] and the block-based weighted average fusion rule in [5].

In the first group, the comparison is performed on a pair of multi-focus images with size of 512×512. We take the classical “Lena” image as a reference image, as shown in Fig. 3(a). We artificially produce a pair of out-of-focus images, by blurring the left part to obtain the image in Fig. 3(b), and then blurring the right part to produce the image in Fig. 3(c). Blurring is accomplished by using a Gaussian low-pass filter. The fusion results using the maximum selection fusion, weighted average fusion and our method are shown in Fig.3 (d), (e) and (f) respectively.

In the second group, multi-modal medical images supplied by Dr. Oliver Rockinger [17] are used as input. The first one is a Computed Tomography (CT) image shown in Fig. 4(a), while the other one is a Magnetic Resonance Image (MRI), see Fig. 4(b). For more information about the images, refer to [17]. The fusion results using the maximum selection fusion, weighted average fusion and our method are shown in Fig.4 (c), (d) and (e) respectively.

It is well known that assessing image fusion performance in a real application is a complicated issue. In many cases qualitative criteria such as visual analysis is used to assess the fusion result. However, a more accurate and reliable evaluation is to combine visual assessment based on a subjective qualitative analysis with a parameter assessment based on an objective/quantitative analysis. Therefore, we firstly evaluate our proposed algorithm perceptually and afterwards, we use several quality measures to compare its results to previous approaches.

4.1 Perceptual quality evaluation

Perceptual evaluation mainly assesses the visual quality of the fused image by means of observation of clarity, contrast and preservation of details.

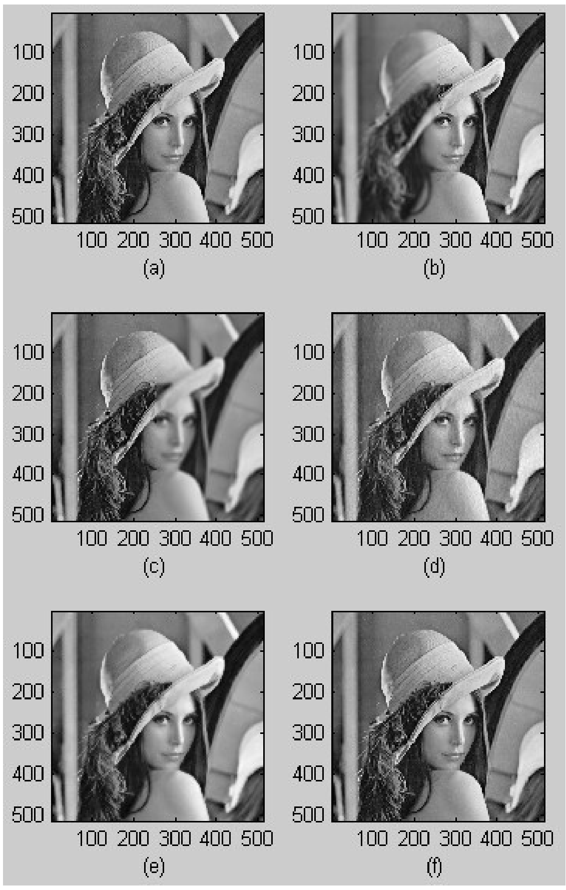


Fig.3 Reference image, multi-focus input images and fused images. (a) Reference image. (b) Focus on the left part. (c) Focus on the right part. (d) Fusion result using maximum selection. (e) Fusion result using weighted average (f) Fusion result using our method.

Based on a visual comparison the fusion results using our method, see Fig. 3(f), contains most of the details of the individual input images in Fig. 3(b) and Fig. 3(c), On the one hand Fig. 3(f) looks smoother that Fig. 3(d), but on the other hand it is clearer than Fig. 3(e). Summarizing, judging the perceptual quality our method performs better than the maximum selection and weighted average methods.

With regard to the visual comparison of the second group, the fusion result using our method in Fig.4 (e) contains more information than the input images in Fig.4 (a) and (b). Fig.4 (e) has more details than Fig.4 (c), whereas Fig.4 (e) has a higher contrast than the image in Fig.4 (d). For a comparison of the image details the enlarged fusion results for all methods are shown in Fig. 5.

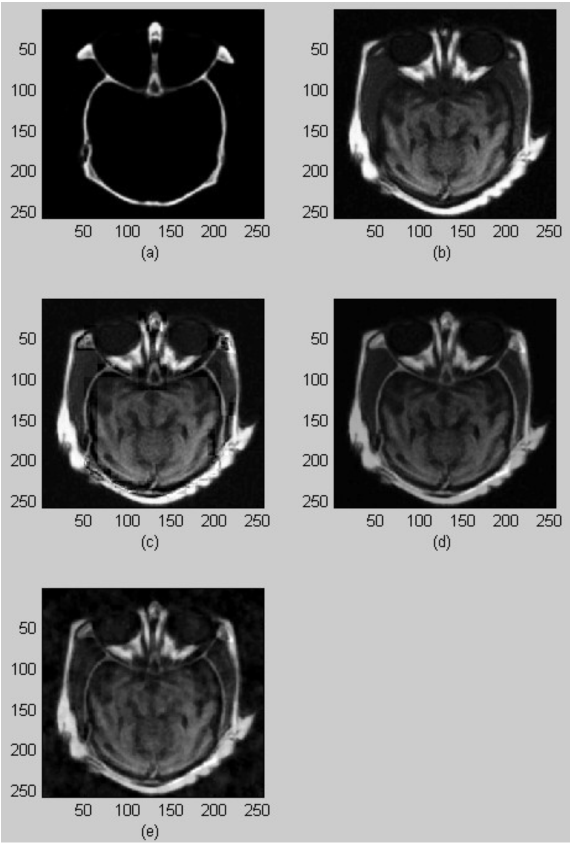


Fig.4 Multi-modal input images and fusion images. (a) CT image. (b) MRI image. (c) Fusion result using maximum selection. (d) Fusion result using weighted average. (e) Fusion result using our method.

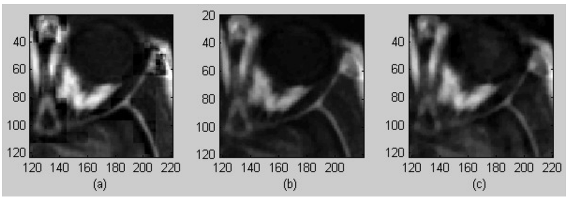


Fig.5 Images in zoom in view. (a) Fusion result using maximum selection. (b) Fusion result using weighted average. (c) Fusion result using our method.

Our approach outperforms the method of maximum selection fusion and weighted average fusion when judging the perceptual quality of the fusion results for both image sets.

4.2 Objective quantity evaluation

In general, there are a few quality measures that are commonly used to evaluate image fusion results: image entropy, mutual information and average gradient.

(a) Image entropy (IE)

Image entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. For an 8-bit single channel image, the image entropy is defined as:

$$H = -\sum_{i=0}^{255} P_i \log_2 P_i \tag{9}$$

where P_i is the probability of gray level i in the evaluated region and it is approximately given by

$$P_i = \frac{f_i}{N} \tag{10}$$

where f_i is the frequency of gray level i and N denotes the total number of pixels in the image. The higher the value of the image entropy is, the more textural information is contained in the (fused) image.

(b) Mutual information (MI)

Mutual Information is often used to evaluate image fusion quality. Let the joint histogram of source image $A(B)$ and the fused image F be $p_{FA}(f,a)(p_{FB}(f,b))$. Then the mutual information between the source image and the fused image is given by

$$I_{FA}(f,a) = \sum_{f,a} p_{FA}(f,a) \log_2 \frac{p_{FA}(f,a)}{p_F(f)p_A(a)} \tag{11}$$

$$I_{FB}(f,b) = \sum_{f,b} p_{FB}(f,b) \log_2 \frac{p_{FA}(f,b)}{p_F(f)p_B(b)} \tag{12}$$

The image fusion performance can be measured by:

$$MI_F^{AB} = I_{FA}(f,a) + I_{FB}(f,b) \tag{13}$$

where larger values imply better image quality.

(c) Average gradient (AG)

The average gradient is a measure of contrast in a photographic image. It is sensitive to reflect the image of the tiny details contrast. It is commonly used to evaluate the clarity of image. We use average gradient as a criterion for image fusion quality. The greater the average gradient value is, the sharper is the image. It can be calculated as:

$$\bar{g} = \frac{1}{n}\sum \sqrt{\frac{(\Delta I_x)^2 + (\Delta I_y)^2}{2}} \tag{14}$$

where n is the size of the image, ΔI_x and ΔI_y are the differences in horizontal and vertical direction respectively.

The performance assessments of the fusion results shown in Figs. 3-4 based on the defined criterions (i.e., IE, MI and AG) are listed in tables 1-2.

Experiment 1

Regarding the "Lena" image, for the purpose of comparing mutual information parameter in detail, we calculate not only the mutual information between the fused image and the individual image, but also the mutual information between the fused image and the original reference image as listed in Table 1. I_{FA} is the mutual information between the fused image and the source image A, while I_{FB} is the mutual information between the fused image and the source image B. MI is the sum of I_{FA} and I_{FB}. I_{FR} is the mutual information between the fused image and reference image. We present here the value using two decimal places due to the limited space in the table.

Table 1 Quantitive evaluation of the multi-focus images shown in Fig.3. ("Lena" image)

Methods	Performance Evaluation Measures					
	IE	I_{FA}	I_{FB}	MI	I_{FR}	AG
Our method	7.12	2.75	2.85	5.60	3.03	2.86
Maximum selection	7.10	2.40	2.57	4.97	2.81	3.59
Weighted average	6.99	2.68	2.71	5.39	3.02	2.21

It is shown in Table.1 that our proposed method outperforms the other methods in terms of IE and MI, which means that the fusion result of our method contains more details than those of the other methods. The visual comparison above also suggests that the fusion result of our method is superior to the result of the maximum selection method and clearer than the result of the weighted average method, though the average gradient value for the maximum selection method is a little bit larger than that for our method. Overall, based on the visual comparison and comparison using objective measures, we can draw the conclusion that our proposed method achieves better performance than the other two methods.

Experiment 2

Regarding the medical image, we only compare the three performance assessment measure (IE, MI and AG), since we do not have the reference image. The results are shown in Table 2.

Table.2 Quantity evaluation of multi-modal images in Fig. 4. (Medical images)

Methods	Performance Evaluation Measures		
	IE	MI	AG
Our Method	6.9763	5.2867	5.1054
Maximum selection	6.6992	5.1544	6.5336
Weighted average	5.8196	3.7439	3.0164

It can be seen easily that our method performs better than the other two methods when comparing the IE and MI results in Table. 2. Taking the visual analysis in paragraph 4.1 into account, we conclude that our method outperforms the methods of maximum selection fusion rule and average gradient fusion rule.

In one word, considering the qualitative analysis and the quantitative evaluation, we conclude that the results of the proposed fusion scheme are superior when compared to the maximum selection fusion rule and the weighted average fusion rule.

5. Conclusions

In the paper, we presented an effective image fusion scheme based on the CS theory. The computational complexity decreases due to the fact that the proposed scheme only needs incomplete measurements rather than acquiring all the samples of the whole image. Moreover, although our method performs the fusion in the sparse domain, it preserves much richer texture information of the individual input images compared with other fusion schemes. Experiments demonstrate the promising performance of our scheme.

6. References

[1] Rick S. Blum, and Zheng Liu, *Multi-Sensor Image Fusion and Its Applications*, Published by: CRC Press. 2005.

[2] D.L.Donoho,"Compressed sensing", IEEE Transactions on Information Theory, 2006, 52(4): 1289-1306.

[3] E.Candes, "Compressive sampling", In: Proceedings of International Congress of Mathematicians. Zurich, Switzerland: European Mathematical Society Publishing House, 2006, 1433-1452.

[4] T. Wan, N. Canagarajah, and A. Achim, "Compressive image fusion", In: Proceeding of 15th IEEE International Conference on Image Processing. San Diego, California, U.S.A: IEEE, 2008, 1308 – 1311.

[5] X.Luo, J.Zhang, J.Yang, and Q.Dai, "Image fusion in compressive sensing", In: Proceeding of 16th IEEE International Conference on Image Processing.Cairo, Egypt: IEEE, 2009, 2205-2208.

[6] Syed Ali Khayam, "The Discrete Cosine Transform (DCT): Theory and Application", *ECE. 802 – 602*: *Information Theory and Coding*, 2003.

[7] E.Candés, J.Romberg, and T.Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information", *IEEE Transactions on Information Theory*. 2006, 52 (2): 489-509.

[8] R.Baraniuk, "Compressive sensing", *IEEE Signal Processing Magazine*, 2007, 24(4): 118-121.

[9] E.Candés and T.Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?", *IEEE Transactions on Information Theory*, 2006, 52(12): 5406-5425.

[10]E.Candés and T.Tao, "Decoding by linear programming", *IEEE Transactions on Information Theory*, 2005, 51(12):4203-4215.

[11] Y.Tsaig and D.L.Donoho, "Extensions of compressed sensing", *Signal Processing*, 2006, 86(3): 549-571.

[12] W. U. Bajwa, J. D. Haupt, G. M. Raz, S. J. Wright, and R. D. Nowak, "Toeplitz-structured compressed sensing matrices", in Proc. IEEE Stat. Sig. Proc. Workshop, Madison, WI, August 2007, pp. 294-298.

[13] E.Candés, J.Romberg and T.Tao, "Stable signal recovery from incomplete and inaccurate measurements", *Communications on Pure and Applied Mathematics*, 2006, 59(8): 1207-1223.

[14] J.Tropp and A.Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit", *IEEE Transactions on Information Theory*, 2007, 53(12): 4655-4666.

[15] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems", *IEEE Journal*, vol. 1, Issue 4, pp.586-597.

[16] D.L.Donoho, "For most large underdetermined systems of linear equations, the minimal l1 norm solution is also the sparsest solution", *Communications on Pure and Applied Mathematics,* 2006, 59(6): 797-829.

[17] http://www.imagefusion.org.

A.2 “Multimodal human detection by sparse feature pursuit”, International Conference on Digital Signal Processing, IEEE, 2011.

MULTIMODAL HUMAN DETECTION BY SPARSE FEATURE PURSUIT

Juanjuan Han, Otmar Loffeld and Klaus Hartmann

Center for Sensor Systems (ZESS)
University of Siegen, Germany

ABSTRACT

Human detection from multimodal image is a challenging task of information extraction and plays a striking important role for the later steps such as classification, recognition, tracking, and so forth. This paper describes an innovative sparse feature-based approach for human detection using the multimodal image. Firstly we consider a human as sparse feature which moves with multimodal image sequences. And afterwards the problem of moving human estimation can be formulated as decomposition of a matrix into a sparse human matrix and a low-rank background matrix. Furthermore, both of the components are exactly recovered by solving convex optimization problem. Finally the sparse feature that contains human is reconstructed to generate the human map. Experimental results on the real multimodal image from a novel 2D/3D vision system verify the effectiveness of our proposed method. Meanwhile the results yield the potential application of matrix decomposition for various multimodal data analysis.

Index Terms—Human detection, multimodal image, sparse feature, matrix decomposition, multi-camera

1. INTRODUCTION

Human detection plays a crucially important role in many application areas, specifically in scenarios involving human motion. Most work has concentrated on the vision system that only operates a visible spectrum camera (e.g. 2D color image) and ignores the other sensor modalities to some extent. In recent few years, Photonic Mixer Devices (PMD) camera has attracted more and more attention due to the fact that it is a powerful device to provide three maps with different modalities (i.e., depth map, modulation amplitude map and intensity image) with help of the Time-of-Flight (ToF) principle [1][3]. The combinational utilization of PMD camera and standard color camera has emerged recently as an unusual potential to spread. To meet the requirement, a new monocular 2D/3D imaging system (i.e., MultiCam [2]) has been developed in our research center.

Due to the fact that human detection in multimodal image plays a crucial role in various application areas, a number of researchers have made effort to obtain good results by applying a variety of methods. The foreground in 2D/3D image is extracted and used for hand tracking as well as gesture recognition by defining a volume of interest in the work of Ghobadi et al. [6]. While Stauffer et al. [9] applied the standard approach of background modeling by Gaussian mixtures and Harville et al. [11] utilized this method to color and depth image. Ghobadi et al. [7] used the combination of edge detection and an unsupervised clustering technique for foreground segmentation. And a rather simple approach to extract foreground from 2D/3D videos which is based on region growing and refrains from modeling the background is evaluated in the work of Bianchi et al. [10]. In the work of Leens et al. [8] the pixel-based background modeling method, called ViBe, is separately applied to the RGB channels of color image and the three channels of PMD camera. The resulting foreground masks are combined via binary image operations.

In this work, we propose a novel approach towards human detection by using sparse feature pursuit for multimodal image data (i.e., simultaneously pursuit sparse feature to represent corresponding moving human for multimodal image). The image with almost stationary background and dynamic human can be considered as the samples of signals that change slowly in time with the sparse feature with arbitrary shape caused by dynamic human. With the sparse prior, we cast a frame of video as a composition of two distinct layers: background and sparse human (i.e., moving foreground). If these frames of video are stacked as the column vectors of a matrix I , the observations matrix I can be expressed as the sum of a background matrix B and a sparse human matrix T which is comprised of sparse feature caused by dynamic human. Due to the background is static or approximate to static, background matrix B therefore exhibits rank-one under the conditions of stationary assumption or low-rank structure. And thus human detection can be transformed into a problem of the pursuit of sparse feature T from the observations $I = B + T$ which consists of low-rank background matrix B and sparse human matrix T . Mathematically speaking, it is mostly possible to be solved by convex optimization. The minimization of a weighted combination of the nuclear norm and the l_1 norm is employed to exactly recover the low-rank matrix B and the sparse feature matrix T [4][5]. The solution is applied to the observations matrix I which represents the image sequences and furthermore the sparse feature that contains human is reconstructed to generate the human map.

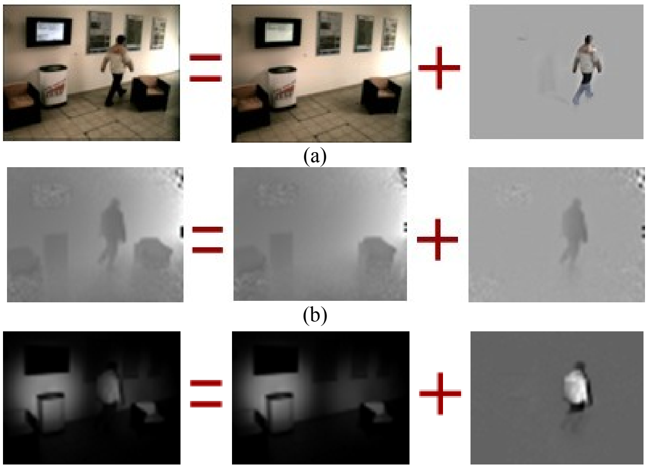
The significant difference from the aforementioned methods is that our proposed approach is based on the underlying assumption of the moving human as the sparse feature. Rather we recover the sparse feature component via convex optimization than try to model the background using the parametric or non-parametric models. It focuses on sparse feature pursuit with arbitrary shape and recovery via matrix decomposition. To our knowledge, matrix decomposition has not yet been applied to multimodal image data. This formulation yields a novel model to detect sparse human for multimodal image data.

The rest of the paper is organized as follows: Section 2 describes the formulation details of our proposed scheme. Sparse matrix recovery and reconstruction of sparse feature map are presented in Section 3. Experimental results on the real multimodal image data validate our proposed method in Section 4. We draw the conclusion in Section 5.

2. PROPOSED SCHEME

2.1. Human as Sparse Feature

From a view of image analysis, each frame of a video sequence consists of two layers: background and foreground. Here we define the background as the static or approximately static region (more or less affected by ambient varying illumination) and the foreground as the region corresponding to the moving human. Therefore each frame can be expressed as the sum of human and background, as expressed in Fig.1.


(a)

(b)

(c)

Fig.1. Image can be regarded as a sum of human and background. (a) Color image. (b) Depth map. (c) Modulation amplitude map.

The moving human moves with the video sequences while the background is mostly static. It is more convenient to consider that the region which the human occupies is clustered into sparse feature of arbitrary shape and the background collection from the video sequence therefore exhibits a low-rank structure. With the prior information about the sparsity of moving human and the low-rank property of background, the image sequences thus are formulated as the sum of low-rank background image sequences structure and sparse foreground image sequences, as described in Fig.2.

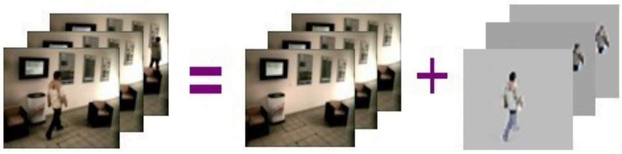


Fig.2. Image sequences can be formulated as a sum of background sequences and image sequences which only contains the moving human.

2.2. Image Data Representation

For the color image, firstly the feature space dimensionality is reduced by using the down-sample technique due to the memory limitation of standard computer. That is, define a color image $\Gamma_{color} \in \mathbb{R}^{h \times w \times 3}$, we obtain $\Gamma_{color} \in \mathbb{R}^{h \times w \times 3} \xrightarrow{down\text{-}sample} Z_{color} \in \mathbb{R}^{(h \times \alpha) \times (w \times \beta) \times 3}$ by performing the down-scale technique on it, where h and w denote the height and width of a color image; α ($0 < \alpha < 1$) and β ($0 < \beta < 1$) are down-sample parameters. And then the color is divided into RGB channels:

(1) Red channel $Z_r = Z_{color}(:,:,R)$

(2) Green channel $Z_g = Z_{color}(:,:,G)$

(3) Blue channel $Z_b = Z_{color}(:,:,B)$)

Finally individual channel is stacked as a column vector to form RGB column vectors, respectively: $Z_* \in \mathbb{R}^{(h \times \alpha) \times (w \times \beta)} \xrightarrow{stack} Z_* \in \mathbb{R}^{d \times 1}$, where $* \in \{r, g, b\}$ and $d = (h \times \alpha) \times (w \times \beta)$.

However, with respect to the low resolution of the depth map (also modulation amplitude map) and furthermore it has the lower feature space dimensionality compared to the color image, what only needs to do are to stack the image data as column vector: $Z_\dagger \in \mathbb{R}^{(h \times \alpha) \times (w \times \beta)} \xrightarrow{stack} Z_\dagger \in \mathbb{R}^{d \times 1}$.

2.3. From Image to Matrix Representation

If the image sequences are stacked as the column vectors to form the matrix I , the matrix B formed by the background image sequences from each images therefore exhibits a low-rank structure and the matrix T formed by the foreground image sequences from each images lays out a sparse structure. And afterwards the image matrix can be formulated as adding the low-rank background matrix with the sparse foreground matrix, as shown in Fig.3.

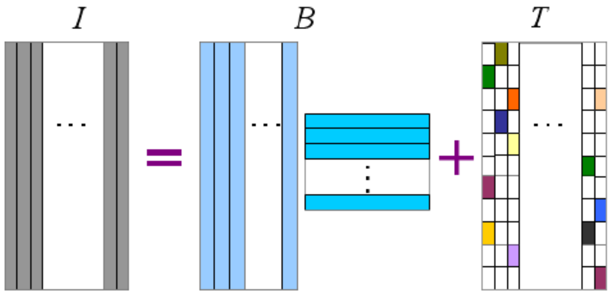


Fig.3. Image matrix can be formulated as a sum of low-rank background matrix and sparse foreground matrix.

2.4. Matrix Decomposition

As already mentioned, it is known that the image matrix is considered as the summation of a low-rank background matrix and a sparse foreground matrix, such a natural attempt is expressed as:

$$I = B + T \qquad (1)$$

where $I = [I_1, I_2, ..., I_n] \in \mathbb{R}^{d \times n}$, $B = [B_1, B_2, ..., B_n] \in \mathbb{R}^{d \times n}$, and $T = [T_1, T_2, ..., T_n] \in \mathbb{R}^{d \times n}$, n is the length of the image sequences. To exactly recovery the two components and furthermore reconstruct the sparse feature for the human, it is more intuitive to consider applying l_0-norm (i.e., the number of non-zero entries) to control the sparsity structure in the matrix and matrix rank to encourage the low-rank structure, that is:

$$\min_{B,T} rank(B) + \lambda \| T \|_0 \quad s.t. \quad I = B + T \qquad (2)$$

where $rank(\cdot)$ denotes the rank of matrix, $\| \cdot \|_0$ is the l_0-norm of matrix, λ is the non-negative balance parameter that trades off the rank of background matrix versus the sparsity of foreground matrix. However, the minimization is not directly tractable due to the fact that the major difficulty on one hand lies in the non-convexity of $rank(B)$ and on the other hand is that it is extremely difficult to minimize the function of l_0-norm. Generally speaking, the decomposition problem of Eq. (2) is NP-hard and there is no effective solution to it. Therefore a computationally tractable alternative, that is, the convex relaxation must be firstly performed on it, as described in the following section.

3. RECONSTRUCTION OF SPARSE FEATURE

3.1. Convex Relaxation

Let the function $f : \mathbb{C} \to \mathbb{R}$, where $\mathbb{C} \subseteq \mathbb{R}^{d \times m}$. The convex hull [14] of f on $\mathbb{C}$ is defined as the largest convex function g so that $g(x) \le f(x)$ for all $x \in \mathbb{C}$. The nuclear norm or the trace norm $\| \cdot \|_*$ has been known as the convex hull of the $rank(\bullet)$ [15]:

$$\| B \|_* \le rank(B), \quad \forall B \in \mathbb{C} = \{B \| \| B \|_2 \le 1\} \qquad (3)$$

And the l_1-norm is the convex envelope of the l_0-norm [14]:

$$\| T \|_1 \le \| T \|_0, \quad \forall T \in \mathbb{C} = \{T \| \| T \|_\infty \le 1\} \qquad (4)$$

Both of the nuclear norm and the l_1-norm functions are convex but non-smooth, and they have exhibited to be effective surrogates of the matrix rank and of the l_0-norm, respectively. Therefore based on the heuristic approximations in Eq. (3) and Eq. (4), we relax the highly non-convex objective function in Eq. (2) by replacing $rank(\cdot)$ with the nuclear norm (i.e., sum of the singular values: $\| \cdot \|_* = \sum_{i=1}^{M} \sigma_i(\cdot)$) and replacing the l_0-norm with l_1-norm (i.e., the sum of the absolute values of matrix entries: $\| \cdot \|_1 = \sum_{ij} | \cdot_{ij} |$), respectively.

And afterwards the relaxation yields a new convex optimization problem: minimization of the nuclear norm and l_1-norm, as shown in Eq. (5). This is the tightest convex relaxation of Eq. (2).

$$\min_{B,T} \| B \|_* + \lambda \| T \|_1 \quad s.t. \quad I = B + T \qquad (5)$$

3.2. Recovery of Sparse Matrix via Convex Optimization

Now the key point is how to solve the convex optimization problem, as expressed in Eq. (5). According to the theoretical conclusion established in [13], the balance parameter λ should be of the order of $\Theta(1/\sqrt{\max\{m,n\}})$. The recent work on convex optimization has yielded few algorithms that solve the relaxed convex problem with a computational cost much lower than that of the classical Principle Component Analysis (PCA). Hereby we apply the method of augmented Lagrange multiplier [12] to solve the problem effectively.

3.3. Human Map Reconstruction from Sparse Matrix

For the color image, each column vector from RGB sparse matrix denotes individual human detection response in the individual RGB channel. The color human map can be reconstructed by combining the human map in three channels.

While for the depth image and modulation amplitude map, the final human detection result is directly generated by reshaping the column vector of the sparse human matrix.

4. EXPERIMENTAL RESULTS

In this section, we verify the effectiveness and evaluate the performance of the proposed method with the experimental results on the two different image database recorded consists of 2D color image and their corresponding depth map and modulation amplitude map that represent different situations for indoor video surveillance. Both of them for evaluation are recorded under normal lighting conditions and bad illumination, respectively. To demonstrate the primary performance of the proposed method, post-processing such as noise removal and connected component analysis are not introduced in this paper.

4.1. Sequence I

The sequence is an indoor scene consists of a university hall, where a man walks from very close to the MultiCam to far away from the MultiCam. There is a TV whose screen is changing mounted on the wall. The ambient lighting conditions are quite stable. And the walking man is well contrasted with the background. The random three frames from this sequence are presented in Fig. 4(a). Due to the memory limitation of the standard computer, the color image with 640×480 pixels is firstly down-sampled to 192×144 pixels using simple down-scale technique. To train these images, sequential 30 frames from the image sequences are used. And the balance parameter is set as $\lambda = 0.8$. It can be seen from the human detection results shown in Fig.4 (b), the walking man is perfectly detected.
In the meanwhile, however, the TV screen and the shadow cast by the walking man also are detected due to the fact that the two regions vary when the man is walking. Since it is not possible to remove the varying background belongs to foreground, we directly apply the proposed method to the original depth image without deleting invalid measurement. Hereby, the balance parameter is fixed as $\lambda = 2$. The original depth images and human detection results from them are shown in Fig.4 (c) and Fig.4 (d), respectively. It can be seen the results in Fig.4 (d) contain much noise where belongs to invalid measurement in modulation amplitude map. If the active modulated infrared light returns from the human point in the scene can not be sensed by PMD sensor, the pixel in the location should be zero in the corresponding modulation amplitude map. After removal of the invalid measurement for the original depth image shown in Fig.4 (e), it is apparent that the proposed method can remove the background details and exactly reconstruct the human, as shown in Fig.4 (f).

Simultaneously apply it to the modulation amplitude map as demonstrated in Fig.4 (g), we set $\lambda = 1$ and obtain the results as shown in Fig.4 (h). Although the original depth map exhibits noisy representation, the extraction of distance information belongs to the region of moving human in depth map is dramatically improved after removal of invalid measurement according to the modulation amplitude map. The combination of the human detection results from 2D color images, 3D depth map and modulation amplitude map can supply big help for post-processing such as human tracking, location and so on that are not introduced in detail in this paper. Therefore, the proposed method can effectively detect the walking person under the normal ambient lighting conditions with multimodal image data for indoor surveillance.

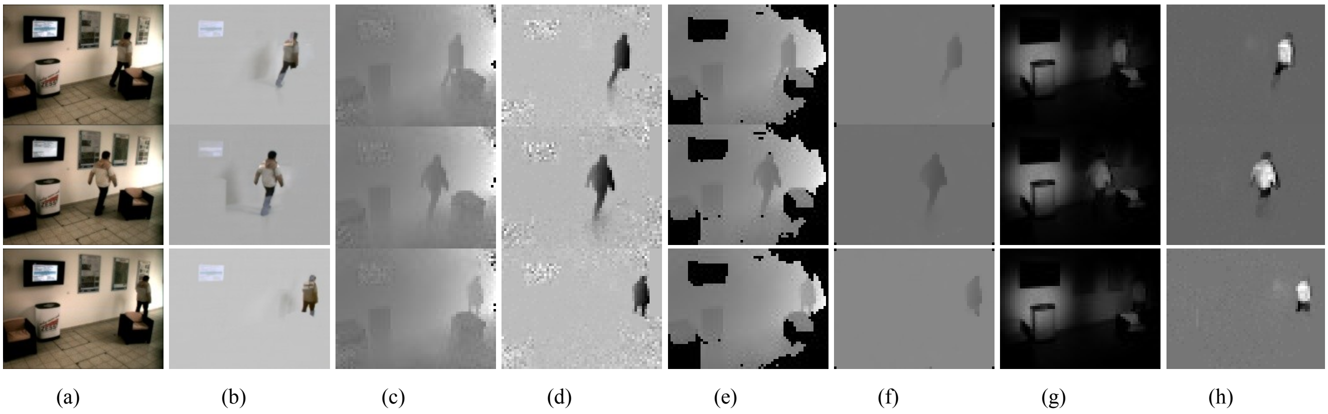


Fig.4. Original images and human detection results. (a) Color images. (b) Human detection results from color images. (c) Original depth images. (d) Human detection results from original depth images. (e) Valid depth images after invalid measurement removal. (f) Human detection results from valid depth images. (g) Modulation amplitude. (h) Human detection results from modulation amplitude.

4.2. Sequence II

For indoor surveillance, it is quite common to have a scene room in a mess under bad illumination. In this image sequence the scene consists of an empty laboratory, where has two sets of fluorescent lights on the ceiling. All the lights are on when outdoor lighting conditions are completely dark, and therefore the illumination could be caliginous and changed gradually. A man walks in and takes an object, finally walks out of the field of view of the MultiCam.

The random four frames from this sequence are presented in Fig. 5. As we mentioned before, firstly we down-sample the color images from 640×480 pixels to 192×144 pixels using simple down-scale technique. And the sequential 30 frames are trained for these multi-modal images. The balance parameter is set as $\lambda=1$.The original color image and human detection results are shown in Fig. 5 (a) and (b), respectively. The human is clearly detected although the environmental illumination is worse than the previous sequence, despite white line or part projected on his body. This is due to the fact that our model learns background template and human motion trajectories captured from the sequential 30 frames, and therefore, different color clusters which allows a quite fair discrimination among colors. However, as we can see from Fig.5 (c), the original depth map delivered from the MultiCam is quite highly noisy in the right part of the image due to invalid measurement in modulation amplitude map. The tradeoff parameter is fixed as λ=1 and the results are presented in Fig. 5 (d). In this case, however, we can observe that the human also is perfectly detected despite of high noise. In the meanwhile the detection results of the modulation amplitude map present a quite promising performance as shown in Fig. 5 (f). Here we also set λ=1 .

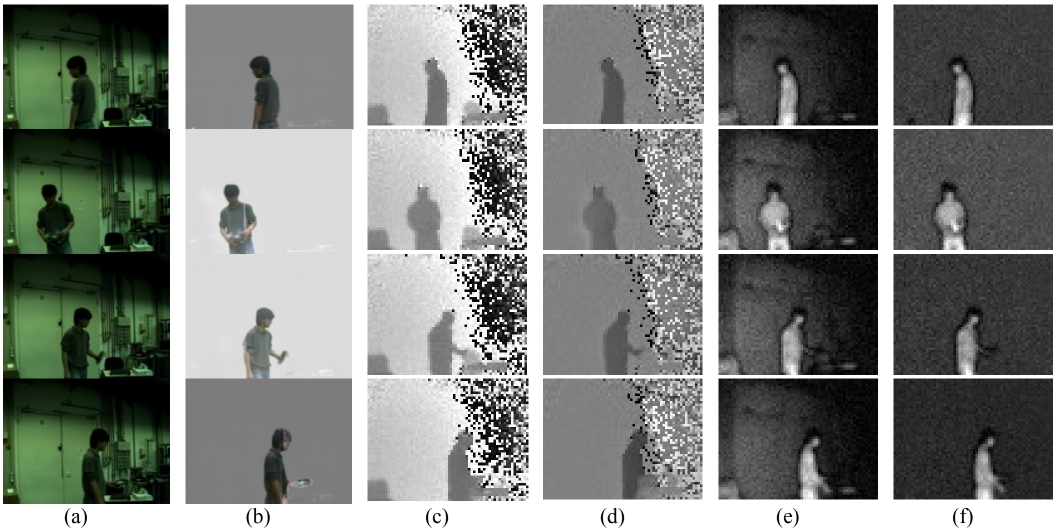


Fig.5. Original images and human detection results. (a) Color image frames. (b) Human detection results from color images. (c) Depth image frames. (d) Human detection results from depth images. (e) Modulation amplitude images. (f) Human detection results from modulation amplitude images.

5. CONCLUSION

The paper presents a novel method towards human detection using sparse feature pursuit for multi-modal images simultaneously provided by a new monocular hybrid 2D/3D imaging system that has been developed based on the ToF principle in recent few years. The procedure of human detection in multi-modal images is implemented based on the following stages. Firstly we make a conception that images with almost stationary background and dynamic human can be considered as the samples of signals that change slowly in time with the sparse feature with arbitrary shape caused by dynamic human. With the sparse prior, we cast a frame of video as a composition of two distinct layers: background and sparse human. And furthermore we recovery the two components based on the recent work on recovery of corrupted sparse low-rank matrix. Finally we reconstruct the human map from the sparse matrix. The experiments on the real image data on the one hand demonstrate the effectiveness of our proposed method, and on the other hand pave a promising way for the combination of multimodal images to yields a greatly improved detection result compared to either type of image data alone.

6. REFERENCES

[1] M. Lindner and A. Kolb, “Lateral and depth calibration of PMD distance sensors,” In Advances in Visual Computing, volume 2, pp.524–533. Springer, 2006.

[2] T. Prasad, K. Hartmann, W. Wolfgang, S. Ghobadi, and A. Sluiter, “First steps in enhancing 3d vision technique using 2d/3d sensors,” in 11. Computer Vision Winter Workshop 2006. University of Siegen: Czech Society for Cybernetics and Informatics, pp. 82–86, 2006.

[3] M. Lindner and A. Kolb, “Calibration of the intensity-related distance error of the PMD TOF-camera,” In SPIE: Intelligent Robots and Computer Vision XXV, volume 6764, pp. 6764–35, 2007.

[4] E. J. Cand`es, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” Technical report, Stanford University, 2009.
http://statistics.stanford.edu/~ckirby/techreports/GEN/2009/2009-13.pdf.

[5] Z. Lin, M. Chen, L. Wu, and Y. Ma, “The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices,” UIUC Technical Report UILU-ENG-09-2215, 2009.

[6] S.E. Ghobadi, O.E. Loepprich, K. Hartmann, and O. Loffeld, “Hand segmentation using 2d/3d images," in IVCNZ 2007 Conference, Hamilton, New Zealand, 2007.

[7] S. Ghobadi, O. Loepprich, O. Lottner, K. Hartmann, O. Loffeld, and W. Weihs. Improved object segmentation based on 2d/3d images. In O. Sablatnig, R.;Scherzer, editor, The Fifth IASTED International Conference on Signal Processing, Pattern Recognition, and Applications (SPPRA 2008), pages 42–47. ACTA Press, 2008.

[8] J. Leens, S. Pi´erard, O. Barnich, M. V. Droogenbroeck, and J.-M.Wagner, “Combining color, depth, and motion for video segmentation,” in ICVS, pp. 104–113, 2009.

[9] C. Stauffer and W. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” IEEE Computer SocietyConference on Computer Vision and Pattern Recognition,vol. 2, pp.252, 1999.

[10] L. Bianchi, P. Dondi, R. Gatti, L. Lombardi, and P. Lombardi, “Evaluation of a foreground segmentation algorithm for 3d camera sensors,” in ICIAP, ser. Lecture Notes in Computer Science, vol. 5716. Springer, pp. 797–806, 2009.

[11] Michael Harville, Gaile Gordon, John Woodfill, "Foreground Segmentation Using Adaptive Mixture Models in Color and Depth," event, pp.3, IEEE Workshop on Detection and Recognition of Events in Video (EVENT'01), 2001

[12] Z. Lin, M. Chen, L. Wu, and Y. Ma, “The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices,” UIUC Technical Report UILU-ENG-09-2215, 2009.

[13] E. J. Cand`es, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? preprint, 2009.

[14] S. Boyd and L. Vandenberghe. Convex Optimization. Cambridge University Press, 2004.

[15] M. Fazel. Matrix Rank Minimization with Applications. PhD thesis, Stanford University, 2002.

A.3 “Moving object extraction using 2D/3D images”, International Conference on Machine Vision, Hong Kong, China, December，2010.

Moving object extraction using 2D/3D images

Juanjuan Han, Klaus Hartmann, Otmar Loffeld and Robert Wang
Center for Sensor Systems (ZESS)
University of Siegen, Germany
{Han, Hartmann, Loffeld, Wang}@zess.uni-siegen.de

***Abstract*—**Moving object extraction has been an active research topic for decades in computer vision for different kinds of applications. This paper presents an innovative method to implement moving object extraction from 2D color image and 3D depth image provided by a novel monocular hybrid 2D/3D imaging system which has been developed based on a promising young technology in recent few years. The problem of moving foreground segmentation in 2D/3D images is formulated as decomposition of a matrix into sparse foreground matrix and low-rank background matrix, and then the two components are exactly recovered via convex optimization afterwards. We further combine the two images of different modalities to surpass the intrinsic limitations of both modalities such as cast shadows from moving object in 2D images and high noise in depth images. Experiments on the real 2D/3D images from this 2D/3D imaging system produce the very promising results and demonstrate the effectiveness of our proposed method.

Keywords-Moving object extraction; background subtraction; Time-of-Flight camera; 2D/3D images; matrix decomposition

I. INTRODUCTION

Detecting and tracking moving objects from a scene is an active and critical element in many applications such as video surveillance and monitoring [1]. When a static camera is used, a common approach to moving object extraction is to segment the foreground region corresponding to the moving objects from the background. Background subtraction technique is a popular approach for foreground segmentation in a still scene [11]. Although the traditional 2D color cameras have got widely used in the past years, there are still some well-known issues need to be specifically addressed in the procedure of background subtraction such as varying illumination, cast shadows, object colors, and so forth.

In recent few years, Photonic Mixer Devices (PMD) camera that can be described as a 3D camera has attracted more and more attention because it can provide distance data between the camera and the points of the scene with help of the Time-of-Flight (ToF) principle [9][10]. Modulated infrared light is emitted by a special lighting device mounted on the camera and reflected on the scene. Afterwards the time denoted as Δt that needed by the infrared light signal to be transmitted and received is measured. This is the reason that this kind of camera is sometimes called ToF camera. Thereby the distance is calculated as $d = c \cdot \Delta t / 2$, where c denotes the speed of the light [7]. Based on this principle, such camera offers a depth image that a color camera can not measure and as a result compensates for the weakness of 2D gray level or color image. However, the technique of ToF range imaging is relatively new and there still exists its own limitations that employed measurement technique is suffering from: statistical noise and low-resolution depth image. So the combinational utilization of PMD camera and standard color camera plays a strikingly important role in many applications. On the one hand a PMD camera offers a depth image which provides the range data information and on the other hand a color camera delivers 2D image with high resolution. We can extract more useful information by fusing the two images of different modalities. For example, a color camera and a 3D depth camera are attached on both sides of a horizontal plane surface and equipped with identical lenses, but such setup comes with the obvious disadvantage (i.e., different view of the two cameras) and thus needs the step of image registration. A new monocular 2D/3D imaging system (i.e., MultiCam [7]) has been developed and it relieves the requirement of image registration due to the fact that the color imager and the PMD sensor share the same lens through an optical splitter.

In this work, we use the MultiCam which on the one hand delivers the high speed range data and on the other hand offers the 2D images at the video frame rate [12], as shown in Fig.1. The resolutions of color imager and PMD sensor are 480×640 pixels and 48×64 pixels, respectively. Meanwhile, we propose a novel method of moving object extraction from 2D/3D images. Assume that a frame in the images can be considered as the sum of static background and foreground such as moving objects. If these frames are stacked as the column vectors of a matrix D, the observations matrix D can be described as the sum of a low-rank background matrix B and a sparse foreground matrix F. Therefore the problem of moving object extraction in 2D/3D images can be considered as a problem of reconstructing a sparse matrix F from the corrupted observations $D = B + F$. Mathematically speaking, the problem can be solved by a convex optimization. The combination of the nuclear norm (also known as the trace norm) minimization and l_1 norm minimization is employed to exactly recover the low-rank matrix B and the sparse matrix F [2][13]. The solution is applied to 2D image and 3D depth image, respectively. Finally the foregrounds in 2D image and 3D depth image are fused in order to improve the performance of moving object extraction and hence foreground separates from background successfully.

This paper is organized as follows: we begin with the related work in section II. Section III describes the details of our proposed method of moving object extraction from 2D/3D images. We present the experimental results and analysis in section IV. And finally we state the conclusions in section V.

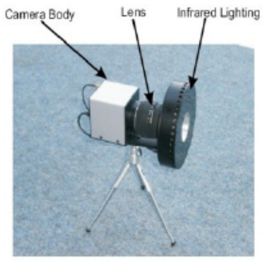


Fig.1. MultiCam

II. RELATED WORK

Due to its important role in various applications, moving foreground extraction in multimodal images has been afforded by several researchers. Stauffer et al. [4] applied the standard approach of background modeling by Gaussian mixtures and Harville et al. [3] used this approach to color and depth images. By defining a volume of interest, the foreground in 2D/3D images is extracted and used for hand tracking as well as gesture recognition in Ghobadi et al. [16]. And in Bianchi et al. [5] a rather simple approach to extract foreground from 2D/3D videos which is based on region growing and refrains form modeling the background is evaluated. In Leens et al. [6] the pixel-based background modeling method, called ViBe, is separately applied to the RGB channels of color images and the three channels of PMD camera. The resulting foreground masks are combined via binary image operations.

Our method is proposed based on the latest research on decomposition of sparse and low-rank matrix and recovery of the two components via convex optimization, and it offers a new approach to extract moving foreground in multimodal images.

III. PROPOSED METHOD

Foreground segmentation or background subtraction is one of the crucial stages in image analysis aiming to isolate the region of the scene corresponding to the moving object. We now describe how to extract the region of moving object from the background in a still scene. Here, we define the foreground as the region corresponding to the moving objects or motion such as people activity, cast shadows from moving object and variability in the background itself opposed to the static background of the video. In this work we consider that an image is composed of the background and the foreground. In a short video recorded from a static camera, the background is almost static and therefore the matrix consists of background is either low-rank structure or well approximately low-rank more or less due to the changing lighting conditions, while the matrix consists of foreground has a sparse structure. This leads to that the problem of moving object extraction can be formulated as a decomposition of a matrix into two components: a low-rank background matrix and a sparse foreground matrix, as well as exactly recovery of the two components. In the following subsections, we describe how to decomposes the matrix and exactly recover the two components of sparse foreground matrix and low-rank background matrix. Then we show how to apply to 2D image and 3D depth image. Finally the foreground activity extraction is done by combining the foreground masks in 2D and 3D images.

A. Decomposition of Sparse and Low-rank Matrix

Assume an image can be expressed as a matrix $I \in \mathbb{R}^{h\times w}$, where h and w are height and width of the image, respectively. If the matrix I is stacked as a column vector $D_i \in \mathbb{R}^M (M = h\times w)$ and an image collection is denoted as $D=[D_1, D_2, ..., D_N] \in \mathbb{R}^{M\times N}$, where N denotes the total number of this image collection. The matrix D therefore can be formulated by adding a background matrix B and a foreground matrix F, that is $D=B+F$. According to the fact that the background matrix B is low-rank and corrupted by some spatial localized foreground activity such as moving objects, we formulate the problem of moving object extraction in 2D and 3D depth images as decomposition of sparse and low-rank matrix. Thanks to the latest research on robust principal component analysis [2], exactly recovering B and F by solving the convex optimization problem:

$$\min_{B,F} \| B \|_* + \lambda \| F \|_1, \quad s.t. \quad D = B + F \tag{1}$$

where $\| \cdot \|_*$ is the nuclear norm of a matrix or sum of the singular values: $\| \cdot \|_* = \sum_{i=1}^{M} \sigma_i(\cdot)$, $\| \cdot \|_1$ represents the l_1 norm or the sum of the absolute values of matrix entries: $\| \cdot \|_1 = \sum_{ij} | \cdot_{ij} |$, and λ is the balance parameter that is set positive.

B. Optimization via Augmented Lagrange Multiplier Method

Now the key point is how to solve the convex optimization problem, as expressed in Eq. (1). Based on the recent work on recovery of corrupted low-rank matrix [13], hereby we apply the method of augmented Lagrange multiplier to solve problem effectively.

The augmented Lagrange function is given by:

$$L_\mu(B,F,Y,\mu) = \\ \|B\|_* + \lambda\|F\|_1 + \langle Y, D-B-F \rangle + \frac{\mu}{2}\|D-B-F\|_F^2 \tag{2}$$

where $Y \in \mathbb{R}^{M\times N}$ is a Lagrange multiplier matrix, μ is a positive scalar, $\langle \cdot,\cdot \rangle$ denotes the matrix inner product, and $\| \cdot \|_F$ denotes the Frobenius norm. The augmented Lagrange multiplier method iteratively estimates both the Lagrange multiplier and the optimal solution by iteratively minimizing the augmented Lagrangian function:

$$(B_{k+1}, F_{k+1}) = \arg\min_{B,F} L_{\mu_k}(B,F,Y_k), \\ Y_{k+1} = Y_k + \mu_k (D - B_{k+1} - F_{k+1}) \tag{3}$$

The iteration definitely converges to the optimal solution of the problem in Eq. (1) [59]. However, it is very difficult to directly solve the first step in the above iteration in Eq. (3). Researchers commonly try to minimize the Lagrangian function approximately against the two unknown variables B and F at one time:

$$B_{k+1} = \arg\min_B L_{\mu_k}(B, F_k, Y_k) \tag{4}$$

$$F_{k+1} = \arg\min_F L_{\mu_k}(B_{k+1}, F, Y_k) \tag{5}$$

For the clearness, the complete procedure to solve the convex optimization based on the method of augmented Lagrange multiplier is summarized in Algorithm 1.

```
Algorithm 1: Optimization via  Augmented Lagrange Multiplier Method
Input:  Observations Matrix $D \in \mathbb{R}^{M \times N}$ , $\lambda$ , $\rho$ , $\mu$ .
while not converged do
    $(U, S, V) = svd(D - F_k + \frac{1}{\mu_k} Y_k)$ ;
    $B_{k+1} = US_{\frac{1}{\mu_k}}[S]V^T$ ;
    $F_{k+1} = S_{\frac{\lambda}{\mu_k}}[D - B_{k+1} + \frac{1}{\mu_k} Y_k]$ ;
    $Y_{k+1} = Y_k + \mu_k (D - B_{k+1} - F_{k+1})$ ;
    $\mu_{k+1} = \rho \mu_k$ ;
    $k = k + 1$ ;
end while
Output:  $B = B_k$ ; $F = F_k$ .
```

where $svd(\cdot)$ denotes the singular value decomposition operator.

C. Implementation Issues

We hereby describe the implementation details for Algorithm 1. Due to the fact that the basic idea of the augmented Lagrange multiplier method is to search for the saddle point of the augmented Lagrange function, rather than directly solving the original constrained optimization problem. We vary the parameter μ starting from the initial value μ_0 and increase it monotonically with each iteration to speedup the convergence until it reaches sufficiently large, in the meanwhile the difference in the value of the cost function is enough small between two consecutive iterations when the iteration definitely converges to the optimal solution of the problem [14].

D. Foreground Segmentation on 2D color Image

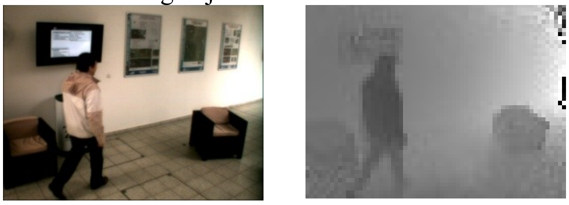
2D color camera usually has high resolution compared to PMD camera. In our work, their resolutions are 480×640 pixels and 48×64 pixels, respectively. To isolate the standard PC memory limitation, we downsample all the test color images to 48×64. After being processed, they will be resized to 480×640 via the upscaling technique. The methods of matrix decomposition and convex optimization above described are sequentially applied to the individual RGB channels of color images, and then the segmentation maps are combined to build an RGB-foreground. With this modality, some regions could be misclassified in the foreground, including the following cases:

- Background is not static.
- There are strong shadows cast by moving objects.
- The color of moving objects is similar to those of the background in a scene.

In addition to the cases above mentioned, other pixels could also be misclassified in the background. For example, the environmental lighting condition is too dark. However, the ToF depth data naturally compensates for these disadvantages and weakness of RGB data. For example, the depth image automatically segments foreground objects from background, a very difficult task in color images. The ToF camera has its own modulated infrared lighting source so that it is almost unaffected by ambient light sources. The detail analysis and discussion are presented in the experiments.

E. Foreground Segmentation on 3D Depth Image

Actually, the PMD camera based on ToF principle also delivers the three channels: distance data, modulation amplitude and intensity image. The distance data is encoded as a depth image, as shown in Fig.2 (b). Fig.2 (a) is the color image corresponding to the depth image. Here we only use the depth information due to the fact that PMD camera uses its own modulated infrared light source and can even works in the complete darkness. Furthermore, the depth image does not suffer from the shadow of moving object.



(a) (b)

Fig.2 (a) Color image. (b) Depth image.

However, it is necessary to perform the pre-processing techniques before applying the complex processing algorithms to the depth images. As we can observe from Fig.2 (b), the distance images provided by the PMD camera are noisy, especially in the corner. To improve the depth image, a median filter is used to filter the statistical noise and smooth the distance data. Owing to the fact that the lighting system does not illuminate the whole field of view of the camera, some pixels in the corner of the depth image take the out of range value and therefore are filtered in order not to affect the accuracy of the post-processing [15]. As the colors have different reflection factors, to some extent the depth image is affected by the color of the object (i.e., the region belongs to the Television in the image) [16]. Some pixels could be classified in the background if the object lies at the same distance from the camera as the background (i.e., the posters mounted on the wall). We simultaneously apply our proposed methods to the depth image and obtain the results, as shown in Fig.3 (b). In order to combine with color images, we upscale the depth image with low resolution to 480×640 by interpolation operation.

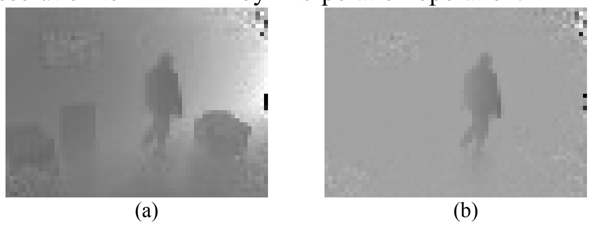


Fig.3 (a) Original depth image. (b) Moving foreground extraction.

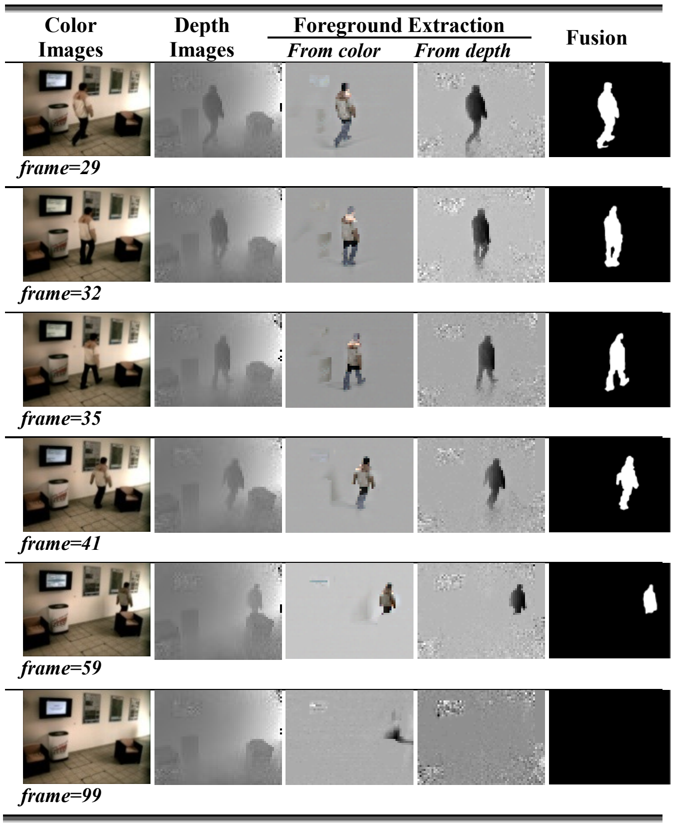
Combining moving foreground from 2D and 3D Images

From the above discussions in section III-D and section III-E, there exists the necessity to combine both images of different modalities to overcome their intrinsic limitations and thereby improve the performance of moving foreground extraction.

IV. EXPERIMENTS

In this experiment, our proposed methods are applied to the real images recorded from this 2D/3D imaging system at the modulation frequency of 20 MHz and the maximum distance for the target of 7.5 meters in an indoor scene. The sequence contains 100 frames of color images with resolution of 480×640 pixels and depth images with resolution of 48×64 pixels. The indoor scene consists of a university hall, where a man comes in and walks away until disappears out of the field of view of the cameras. The ambient lighting conditions are very stable and moving object is well contrasted with the background (background is mostly static). Due to the limitation of standard computer memory, we downsample the color image to 48×64 pixels and process them sequentially by individual RGB channels. After completing processing them, we upscale the results to 480×640 via interpolation operation. The parameters λ and ρ are fixed as $\lambda=1$ and ρ=1.25, respectively. μ is initialized as $\mu_0 = 1/norm(D)$, where $norm(D)$ denotes the norm of observations D. And afterwards the results of the three channels are merged as RGB-foreground, as shown in the third column of Table 1. As we can see from the first column of Table 1, the moving object has been exactly extracted from the original color images. However, the strong shadows cast by moving object can also be observed in the color sequence. Especially, the shadow of moving object still is left in the color image (frame=99) when the man has already come out, as shown in the last row of the third column in Table 1. To solve this problem, the depth images are also processed to compensate for the drawbacks in the color images. The filter is firstly used since the noise in the top-right corner of the depth images is extremely heavy as we can see from the second column in Table 1. The processed results are presented in the forth column in Table 1. The moving object without the shadow is also completely detected from the depth images. By combining them, we obtain the final moving foreground extraction results, as shown in the fifth column in Table 1. From such results, we can observe that the moving object is successfully extracted from 2D/3D images.

TABLE 1. EXPERIMENTAL RESULTS



V. CONCLUSTIONS

The paper presents a novel method of moving object extraction using 2D color and 3D depth images provided by a new monocular hybrid 2D/3D imaging system that has been developed based on the ToF principle in recent few years. The procedure of background subtraction in 2D and depth images is implemented based on the recent work on recovery of corrupted low-rank matrix. And then we are able to extract moving foreground by combining the depth images with those of RGB camera. The experiments demonstrate the effectiveness of our proposed method and provide promising results.

REFERENCES

[1] M. Piccardi, “Background subtraction techniques: a review,” in Proc. of IEEE SMC 2004 International Conference on Systems, Man and Cybernetics, The Hague, The Netherlands, October 2004.

[2] E. J. Cand`es, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” Technical report, Stanford University, 2009. http://statistics.stanford.edu/~ckirby/techreports/GEN/2009/2009-13.pdf.

[3] M. Harville, G. Gordon, and J. Woodfill, “Foreground segmentation using adaptive mixture models in color and depth,” Detection and Recognition of Events in Video, IEEE Workshop on, vol. 0, pp.3, 2001.

[4] C. Stauffer and W. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” IEEE Computer SocietyConference on Computer Vision and Pattern Recognition,vol. 2, pp.252, 1999.

[5] L. Bianchi, P. Dondi, R. Gatti, L. Lombardi, and P. Lombardi, “Evaluation of a foreground segmentation algorithm for 3d camera sensors,” in ICIAP, ser. Lecture Notes in Computer Science, vol. 5716. Springer, pp. 797–806, 2009.

[6] J. Leens, S. Pi´erard, O. Barnich, M. V. Droogenbroeck, and J.-M. Wagner, “Combining color, depth, and motion for video segmentation,” in ICVS, pp. 104–113, 2009.

[7] T. Prasad, K. Hartmann, W. Wolfgang, S. Ghobadi, and A. Sluiter, “First steps in enhancing 3d vision technique using 2d/3d sensors,” in 11. Computer Vision Winter Workshop 2006. University of Siegen: Czech Society for Cybernetics and Informatics, pp. 82–86, 2006.

[8] Gvili, R., Kaplan, A., Ofek, E., Yahav, G.: Depth Key. SPIE Electronic Imaging (2006).

[9] M. Lindner and A. Kolb,“Lateral and depth calibration of PMD distance sensors,” In Advances in Visual Computing, volume 2, pp. 524–533. Springer, 2006.

[10] M. Lindner and A. Kolb, “Calibration of the intensity-related distance error of the PMD TOF-camera,” In SPIE: Intelligent Robots and Computer Vision XXV, volume 6764, pp. 6764– 35, 2007.

[11] L. Maddalena and A. Petrosino, “A Self-Organizing Approach to Background Subtraction for Visual Surveillance Applications,” IEEE Transactions on Image Processing, DOI 10.1109/TIP.2008.924285, Vol. 17, no. 7, pp. 1168-1177, July 2008.

[12] S.E.Ghobadi, O.E. Loepprich, F. Ahmadov, J.Bernshausen, K. Hartmann, and O.Loffeld, “Real time hand based robot control using 2D/3D images,” In International Symposium Visual Computing (ISVC), vol. 5359 of LNCS, Springer, pp. 307–316, 2008.

[13] Z. Lin, M. Chen, L. Wu, and Y. Ma, “The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices,” UIUC Technical Report UILU-ENG-09-2215, 2009.

[14] D. P. Bertsekas, Nonlinear Programming. Athena Scientific, 2004.

[15] S.E. Ghobadi, O.E. Loepprich, O. Lottner, W. Weihs, K. Hartmann, and O. Loffeld, “Analysis of the Personnel Safety in a Man-Machine Cooperation Using 2D/3D Images,” RISE, Spain, 2008.

[16] S.E. Ghobadi, O.E. Loepprich, K. Hartmann, and O. Loffeld, “Hand segmentation using 2d/3d images," in IVCNZ 2007 Conference, Hamilton, New Zealand, 2007.

A.4 “Depth map resolution enhancement for 2D/3D imaging system via compressive sensing”, International Symposium on Photoelectronic Detection and Imaging, SPIE, 2011.

Depth map resolution enhancement for 2D/3D imaging system via compressive sensing

Juanjuan Han*, Otmar Loffeld, Klaus Hartmann
Center for Sensor Systems (ZESS), University of Siegen, Germany

ABSTRACT

This paper introduces a novel approach for post-processing of depth map which enhances the depth map resolution in order to achieve visually pleasing 3D models from a new monocular 2D/3D imaging system consists of a Photonic mixer device (PMD) range camera and a standard color camera. The proposed method adopts the revolutionary inversion theory framework called Compressive Sensing (CS). The depth map of low resolution is considered as the result of applying blurring and down-sampling techniques to that of high-resolution. Based on the underlying assumption that the high-resolution depth map is compressible in frequency domain and recent theoretical work on CS, the high-resolution version can be estimated and furthermore reconstructed via solving non-linear optimization problem. And therefore the improved depth map reconstruction provides a useful help to build an improved 3D model of a scene. The experimental results on the real data are presented. In the meanwhile the proposed scheme opens new possibilities to apply CS to a multitude of potential applications on various multimodal data analysis and processing.

Keywords: Super-resolution, compressive sensing, Time of Flight camera, depth map, resolution enhancement

1. INTRODUCTION

The ability to capture 3D model of a scene has attracted more and more attentions. However, to build 3D geometric information of real environments in an automated and fast way is still a challenge task. The conventional image sensors such as CCD/CMOS measure the color image information of the scene, but it lacks the depth information of the scene. Even for static scenes there is no low-priced off-the shelf system available, which provides full-range range information in real-time and low-cost way. Laser scanning techniques merely sample a scene row by row with a single laser device, are rather time-consuming and impracticable for dynamic scenes. Stereo vision camera systems suffer from the inability to match corresponding points in homogeneous regions [3].

As a recent development in imaging hardware, the Time-of-Flight (ToF) cameras has been introduced that capture 3D depth map by measuring the return travel time of a modulated infrared light wave-front emitted from the lighting system mounted on the sensor. The combinational utilization of ToF camera and standard color camera has emerged recently as an unusual potential to spread due to the fact that on one hand it can provide color image and on the other hand it delivers depth information between the camera and the object in a scene. Therefore to meet the requirement, a new monocular 2D/3D imaging system (i.e., MultiCam [2]) has been developed in our research center.

Visual input from the color camera delivers high-resolution texture image but meanwhile increases the requirement of enhancement of the depth map calculated from the distance information output of ToF camera in term of spatial resolution. Unfortunately, the depth map suffers from the drawback of far too low resolution due to the restriction of the range sensor. To address this problem, a lot of interesting works have been done to enhance the spatial resolution of depth map. Prasad et al. 0 applied the interpolation method into the depth map. Sebastian et al. proposed an approach that used several depth maps for the super-resolution reconstruction of a depth map [5][8]. Bilateral filtering of the cost volume is used in [4] and the Joint Bilateral Up-sampling method is proposed in the work of [6]. Rajagopalan et al. [10] proposed a Markov-Random-Field (MRF) based resolution enhancement method from a set of low-resolution depth recordings that formulates the up-sampled 3D geometry as the most likely surface given several low resolution measurements. And the similar method based on MRF energy minimization framework is proposed in [7].

*han@zess.uni-siegen.de

In this work, to overcome the problem, we propose a novel approach to enhance the resolution of depth image by adopting the recent emerging theory framework of CS. The problem of estimating and reconstructing high-resolution depth map from low-resolution depth observations is basically ill-posed problem and prior constraints must be imposed to enforce regularization. We build the super-resolution model in the CS framework way. The desired super-resolution (SR) depth image is estimated and furthermore reconstructed by solving non-linear optimization problem. Therefore the improved range image reconstruction provides a useful help to build an improved 3D model of a scene.

The remainder of this paper is organized as follows. We briefly introduce the 2D/3D imaging system in Section 2. Section 3 describes the problem model for super-resolution and the compressive sensing model. We present an approach to solution for the super-resolution problem using the compressive sensing framework. And Section 4 demonstrates the experimental results. Finally we draw a conclusion and make a discussion in Section 5.

2. 2D/3D IMAGING SYSTEM

The multimodal data acquisition device used in our setup is a new monocular 2D/3D imaging system, as shown in Fig.1. This vision system is comprised of two imaging sensors: a standard 2D sensor (CCD/CMOS) and a Photonic Mixer Device (PMD) range sensor. The PMD is an implementation of an optical Time-of-Flight sensor, able to capture distance data between the sensor and the object points of a scene at a video frame rate [9]. The depth image is obtained by coding the distance data afterwards. However, current PMD range sensors suffer from the drawback of noise and low-resolution, typically 64×48 up to 204×204 pixels, which is small compared to standard RGB sensors.

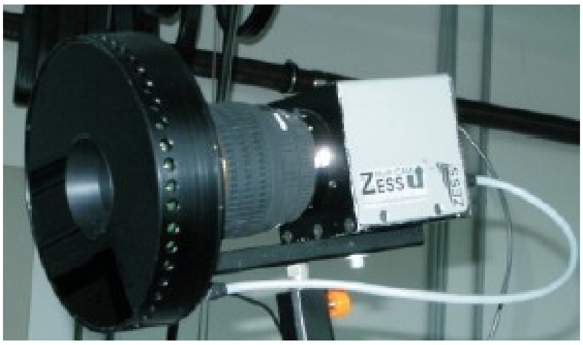


Figure 10. The 2D/3D imaging system developed in our research center.

3. SIGNAL MODAL

3.1 Problem modal

Super-resolution (SR) is a technique that offers the promise of overcoming some of the inherent resolution limitations of low-cost imaging system. SR algorithms attempt to generate a single high-resolution (HR) image from one or more low-resolution (LR) images of the same scene. The goal is to reconstruct the high-frequency missing information in one way that approximates the desired HR image as closely as possible. There are both single-image and multiple-images variants of SR. Multiple-images based SR algorithms utilize the sub-pixel shifts between multiple low-resolution images of the same scene [57]. They create an improved resolution image fusing information from all LR images [13]. However, how to recover missing information from a single LR image is more interesting and challenging. That is to say, the problem of single-image SR is particularly important because our application in which only a single, LR depth map is available and the up-sampling must be applied as a post-processing procedure. It is our goal to obtain high-resolution depth map of a static scene despite the significant noise in the raw data. We enhance X-Y measurement resolution and meanwhile reduce the overall random noise level by performing SR technique.

Let us assume we have a LR depth image, denoted as y . It can be formulated to be obtained through the observations given by the model:

$$y = DGf \tag{1}$$

where D and G respectively stand for the down-sampling (i.e., decimation) and blur operators, and f is the desired HR image of the scene subject to reconstruction. The problem model can be expressed as Fig.2.

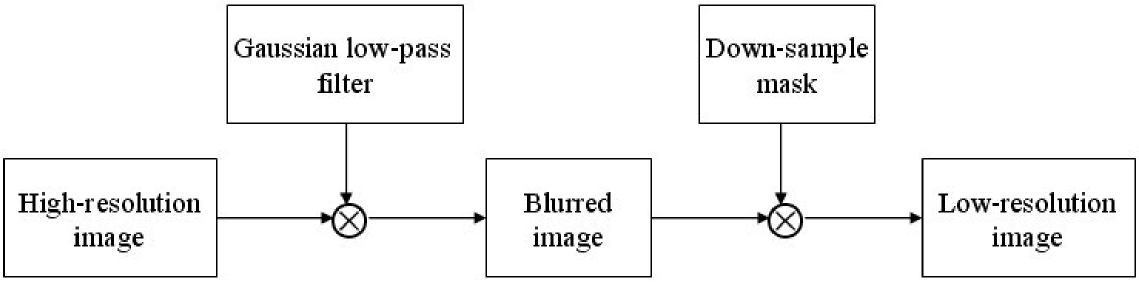


Figure 2. Problem model.

Mathematically speaking, the SR issue is an inverse problem. However, this issue is also an extremely ill-posed problem since there exist infinitely many HR images f meet the reconstruction constraint (i.e., the formulation of Eq. (1)) for a given LR y. Fortunately, the recently emerging inversion theory of CS provides us a great tool for solving the inverse problem. In the next two sub-sections, we describe the formulation of CS theory and problem modeling with compressive sensing framework, respectively.

3.2 Compressive sensing modal

The theory of compressive sensing states that, a signal can be exactly recovered from a small number of random linear measurements if it is sparse in some basis through non-linear optimization [11][12].

Sparsity and random measurement: In a typical framework of compressive sensing, a signal vector $f_0 \in \Re^N$ can be represented by the form $f_0 = \Psi x_0$, where $\Psi \in \Re^{N\times N}$ denotes an orthonormal basis, and $x_0 \in \Re^N$ satisfies $\| x_0 \|_0 = K << N$ ($\|\cdot\|_0$ is l_0 norm.). Due to the sparsity of f_0 relative to the basis Ψ, it is not necessary to sample all N values of f_0. Instead, the CS theory establishes that f_0 can be recovered from a small number of projections onto an incoherent set of measurement observations [11][12]. To measure f_0, we compute $M << N$ linear projections of f_0 via the matrix-vector multiplication

$$y = \Phi f_0 = \Phi\Psi x \tag{2}$$

where $\Phi \in \Re^{M\times N}$ is the measurement matrix.

Restricted Isometry Property (RIP): CS addresses the problem of solving for f_0 when $M << N$, i.e., Θ=$\Phi\Psi$ is severely underdetermined. This is an ill-posed problem as there are an infinite number of candidate solutions for f_0. Nevertheless, the sparse signal f_0 can be accurately estimated if the measurement matrix Φ in conjunction with Ψ satisfies a technical condition called Restricted Isometry Property [12]. That is to say, Θ meets the RIP of order s if there exists a constant $\delta_s \in (0,1)$ for which

$$(1-\delta_s) \| v \|_2 \le \| \Theta v \|_2 \le (1+\delta_s) \| v \|_2 \tag{3}$$

holds for all s-sparse $v \in \Re^N$. As a matter of fact, the RIP presents that a measurement matrix will be valid if every possible set of v columns of Θ forms an approximate orthogonal set. The examples of matrices that have been proven to satisfy the RIP include independent and identically distributed (iid) Gaussian random matrices, Bernoulli matrices, and partial Fourier matrices [13]. An alternative approach to stability is to ensure that the measurement matrix Φ is incoherent with the sparsifying basis Ψ in the sense that the vector $\{\phi_j\}$ cannot sparsely represent the vector $\{\varphi_j\}$ and vice versa [11][12][15].

Reconstruction algorithm: The reconstruction algorithms often rely on an optimization, which searches for the sparsest coefficients x_0 that agree with the measurements y. If M is sufficiently large and x_0 is strictly sparse, x_0 is the solution to the l_0 minimization:

$$\hat{x}_0 = \arg\min \| x \|_0 \quad s.t. \quad y = \Phi\Psi x \tag{4}$$

However, to solve this l_0 minimization is NP-problem [17]. Fortunately, the revelation that supports the CS theory is that a computationally tractable optimization problem yields an equivalent solution. We need to replace the l_0 minimization with l_1 minimization:

$$\hat{x}_1 = \arg\min \| x \|_1 \quad s.t. \quad y = \Phi\Psi x \tag{5}$$

This l_1 optimization problem, also known as Basis Pursuit [18], can be solved by linear programming approaches. However, the l_1 optimization problem requires cubic computation in general and therefore the cubic complexity renders it impractical for many applications. For this reason, a flurry of research on faster algorithms has been motivated. Iterative greedy algorithms such as Orthogonal Matching Pursuit (OMP) [21], Regularized Orthogonal Matching Pursuit (ROMP) [19] and CoSaMP [20] have been investigated. The general framework of CS theory is shown in Fig.3. The high-dimensional signal is transformed to coefficients matrix via sparsifying transform. Then we obtain the low-dimensional observations by applying random measurement to the coefficients matrix. Finally, the original high-dimensional signal can be exactly reconstructed from the small observations by using the recover algorithms.

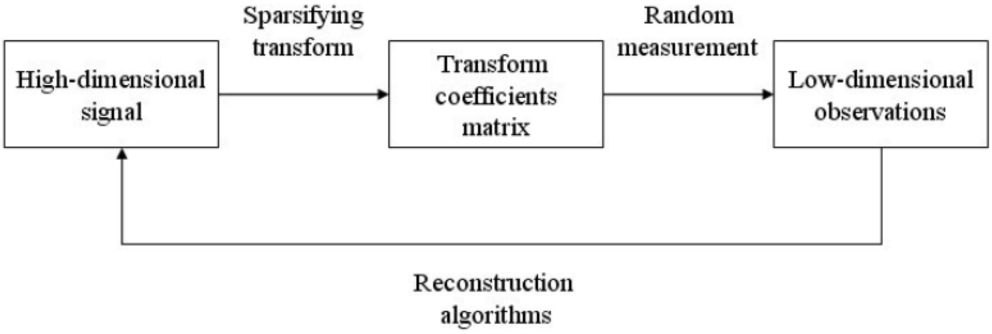

Sparsifying transform
Random measurement
High-dimensional signal
Transform coefficients matrix
Low-dimensional observations
Reconstruction algorithms

Figure 3. The framework of compressive sensing

3.3 Super-resolution modeling via compressive sensing framework

By comparing the problem model with the CS model, as described in Eq. (1) and Eq. (2), we observe that the similarity of them is the inverse and ill-posed problem (i.e., Both of them want to estimate the high dimensional signal from the low dimensional signal). Initially, it seems not possible since the M samples of y yield a $N-M$ dimensional subspace of possible solutions for the original f that would match the given observations. However, a key assumption of CS framework offers a crack to this non-possibility, i.e., the transform version of the signal f is sparse in some basis Ψ. It is not an unreasonable assumption since the depth map is a digital image by coding the distance between the sensor and the object point of the scene. Therefore we consider using the wavelet basis as the sparsifying basis Ψ to sparsely present the depth map, and here $f = f_0$. In order to satisfy the RIP as described in Eq. (3) with high probability, the measurement matrix Φ is modeled as

$$\Phi = DG \tag{6}$$

where D is the down-sampling mask and G denotes the Gaussian filter mask. In this work, we consider the Gaussian filter as a multiplication of a Gaussian function in the Fourier domain:

$$G = F^{-1} G_{*} F \tag{7}$$

where F is the 2D Fourier transform matrix, G_{*} denotes the diagonal matrix whose diagonal elements correspond to the Gaussian function and F^{-1} is the inverse Fourier transform matrix. Therefore the measurement matrix can be expressed as $\Phi = DF^{-1}G_{*}F$. And the super-resolution problem can be modeled as Eq. (8) with the framework of CS.

$$y = DF^{-1} G_{*} F \Psi x \tag{8}$$

With this formulation in hand, to reconstruct the HR image, we apply the ROMP algorithm to solve the optimization problem. Given a LR depth image y , we obtain the wavelet transform x using ROMP algorithm. And then we apply the inverse wavelet transform to x , afterwards reconstruct the desired high-resolution depth map f .

4. EXPERIMENTAL RESULTS

In this section, we demonstrate the experimental results from two image samples to evaluate the performance of the proposed method. For the first test sample, firstly, we directly apply the Gaussian low-pass filter operator and decimation operator to the original HR image in order to obtain the LR image. And then we adopt the proposed approach on the observed LR image to reconstruct the desired HR image. The reconstructed HR image is compared with the original HR image using the compared methods include perceptual quality evaluation and quantitative evaluation. For the second test sample, we would like to directly apply the proposed method to obtain the HR depth map, since only with the LR depth image as the measurement observations in hand instead of HR depth image.

4.1 Test I

Since the real depth map suffers from both of drawbacks of lower resolution and high noise, in the first test sample we use a standard grayscale image which often is used in image processing community to evaluate the performance of the proposed method.

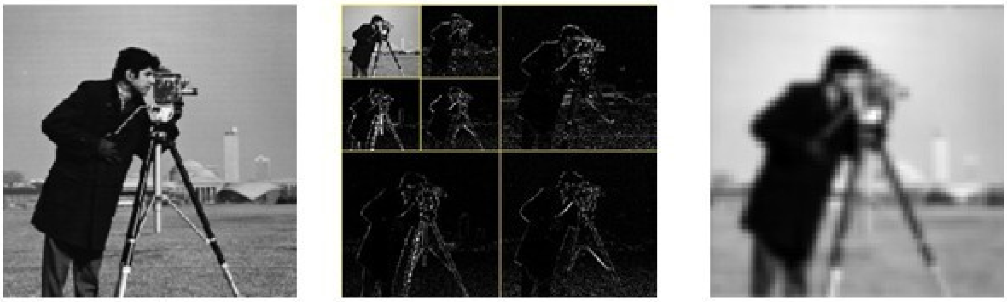


Figure 3. (a) The original HR image. (b) The corresponding wavelet transform with 'haar' at level 2. (c) The LR image after applying Gaussian low-pass filter and decimation operators.

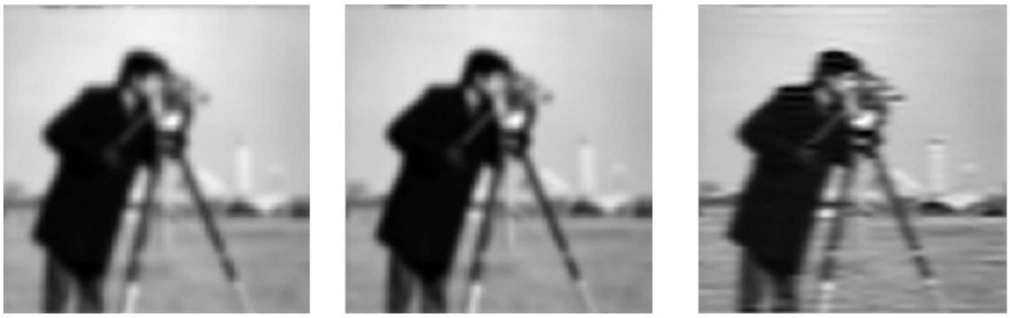


Figure 4. (a) SR result from Fig.3 (c) via 'bilinear' interpolation. (b) SR result from Fig.3 (c) via 'bicubic' interpolation. (c)SR result Fig.3 (c) via the proposed method.

We firstly use the wavelet transform with 'haar' at level 2 to analyze the sparsity of the original image, as shown in Fig.3 (b). The original image (as shown in Fig.3 (a)) exhibits the highly sparsity under the wavelet basis. With the sparsity property in hand, we then blur it with the Gaussian low-pass filter and obtain a sub-sample version with down-sample mask, as shown in Fig.3 (c). To test the propose method, we up-scale the image using various algorithms. In Fig.4 we compare the results of our method against the standard approaches: bilinear interpolation and bicubic interpolation. From a point of view of perception, the result of our method is better than the other approaches with sharper details. For example, one can observe the cameraman's eye in the image reconstructed by the proposed approach. In the meanwhile, we evaluate the results with Peak Signal-to-Noise Ratio (PSNR), the most widely used objective image quality metric. However, an interesting result is that the PSNR value from the proposed method is lower than the other methods. As we have known, the PSNR value does not perfectly correlate with a perceived visual quality due to the non-linear behavior of the human visual system.

4.2 Test II

The depth map (i.e., Fig.5 (a)) used in this work was 204×204 pixels in size with a bit depth of 8 bits. It was captured from the PMD camera of MultiCam monitoring a natural scene with a man walked in the field of view of the camera. However, a big difference from the first test sample is that we directly apply the proposed method to the depth map instead of the down-sample version of HR image due to the fact that only LR depth map is available. Firstly we analyze the sparsity of the original LR depth map in wavelet domain. As we can observe from Fig. 5 (b), it can be sparsely represented in wavelet transform with 'db2' at level 2. With this precondition in hand, we assume that the desired HR depth map is also sparse in the wavelet domain. Therefore we invoke the super-resolution method via compressive sensing framework and apply the Gaussian low-pass filter and point down-sample mask to the desired HR depth map. With y as the given LR depth map, we compute the sparse coefficients under the wavelet basis x and afterwards reconstruct the desired HR depth image using the equation of $f = \Psi x$.

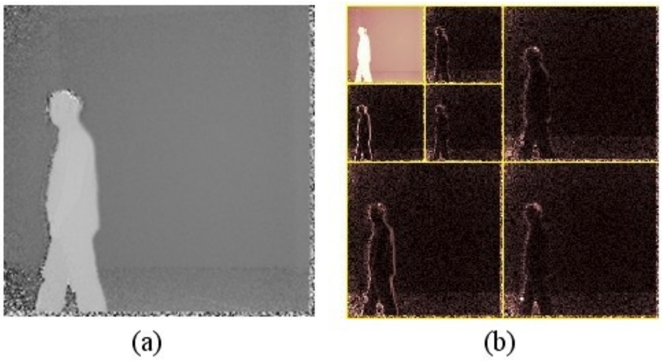


Figure 5. (a) Original LR depth map. (b) The corresponding wavelet transform with 'db2' at level 2.

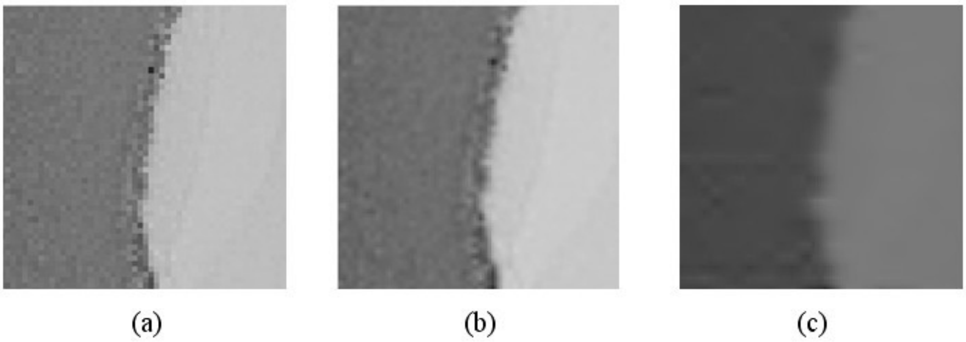


Figure 6. (a) Part of Original LR depth map. (b) SR result via 'bilinear' interpolation. (c) SR result via the proposed method.

For the clarity, we use a part of the results from interpolation and the proposed methods. The part of the original LR depth map is shown in Fig.6 (a). The results are shown in Fig. 6 (b) and (c), respectively. As we can see, the proposed method produces the smoother and clearer edge than the other method. And meanwhile we should also note that the proposed method not only enhances the edge, but also implements the function of noise removal to some extent. Therefore, the proposed method produces a better performance.

5. CONCLUSTIONS AND DISSCUSSIONS

In this work, we firstly study the problem model of super-resolution and compressive sensing theory model. We build the super-resolution signal model via the framework of compressive sensing by the comparison and observation of the both models. The depth map is sparsely represented under the wavelet basis. As shown in the results, the 'haar' wavelets at level 2 we used for compression worked well in the standard gray-level image, while the 'db' wavelets at level 2 worked well in the depth image. We suppose that there might be better-suited wavelet basis for the sparsifying basis. We design the measurement matrix using the multiplication of a point down-sample mask and a Gaussian low-pass filter. The measurement matrix is sensitive to the choice of the sparsifying basis according to the RIP, which means there seems to exist better sparsifying basis such as complex wavelets for sparse representation. In the meanwhile, the random point down-sample mask can be used to improve the effectiveness of the proposed method. In addition, the reconstruction algorithms also play an important role for the result, therefore the other recover algorithms would be interesting to explore.

REFERENCES

[1] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," Image processing, IEEE Transactions on, vol. 13, no. 10, pp. 1327–1344, Oct. 2004.

[2] T. Prasad, K. Hartmann, W. Wolfgang, S. Ghobadi, and A. Sluiter, "First steps in enhancing 3d vision technique using 2d/3d sensors," in 11st Computer Vision Winter Workshop 2006. University of Siegen: Czech Society for Cybernetics and Informatics, pp. 82–86, 2006.

[3] Andreas Kolb, Erhardt Barth, and Reinhard Koch, "ToF Sensors: New Dimensions for Realism and Interactivity, " In CVPR 2008 Workshop on Time-of-Flight-based Computer Vision (TOF-CV), 2008.

[4] Q. Yang, R. Yang, J. Davis and D. Nister, "Spatial-depth super resolution for range images," Proc. IEEE. Conf. Comp. Vision and Pattern Recogn., 2007.

[5] S. Schuon, C. Theobalt, J. Davis and S. Thrun, "High-quality scanning using time-of-flight depth superresolution," Proc. IEEE Conf. Comp.Vision and Pattern Recogn. Workshops, Jun. 2008.

[6] J. Kopf, M. F. Cohen, D. Lischinski and M. Uyttendaele, "Joint bilateral upsampling," ACM Trans. Graphics, vol. 26, no. 3, article. 96, Jul. 2007.

[7] Huhle, B., Fleck, S., and Schilling, A., "Integrating 3D time-of-flight camera data and high resolution mages for 3DTV applications," Proc. 3DTV-Conference, pp. 1–4 (2008)

[8] S. Borman and R. L. Stevenson, "Super-resolution from image sequences – a review," Proc. Midwest Symp. Circuits and Systems, vol.5, pp. 374-378, Apr. 1998.

[9] Ghobadi, S., Loepprich, O., Lottner, O., Hartmann, K., Loffeld, O., and Weihs, W., "Improved Object Segmentation Based on 2D/3D Images," The Fifth IASTED International Conference on Signal Processing, Pattern Recognition, and Applications (SPPRA 2008), Innsbruck, Austria, 13.-15. February, 2008, pp. 42-47.

[10] A. Rajagopalan, A. Bhavsar, F. Wallhoff, and G. Rigoll, "Resolution Enhancement of PMD Range Maps, " Lecture Notes in Computer Science, 5096:304–313, 2008.

[11] E. J. Cand`es, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, " IEEE Trans. on Info. Theory, vol. 52, no. 2, pp. 489–509, 2006.

[12] D. L. Donoho, "Compressed sensing, " IEEE Transactions on Information Theory, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[13] http://en.wikipedia.org/wiki/Super-resolution

[14] E. J. Cand`es and T. Tao, "Decoding by linear programming," IEEE Trans. Inform. Theory 15(12), pp. 4203–4215, 2005.

[15] J. Tropp and A. C. Gilbert, "Signal recovery from partial information via orthogonal matching pursuit," April 2005, www-personal.umich.edu/_jtropp/papers/TG05-Signal-Recovery.pdf.

[16] E. J. Cand`es and T. Tao, "Near optimal signal recovery from random projections: universal encoding strategies?" IEEE Transactions on Information Theory, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[17] E. Candès, M. Rudelson, T. Tao, and R. Vershynin, "Error correction via linear programming," Found. Comput. Math., pp. 295–308, 2005.

[18] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," SIAM J. Sci. Comput., vol. 20, no. 1, pp. 33–61, 1998.

[19] D. Needell and R. Vershynin, "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit," 2007, preprint.

[20] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," Appl. Comput. Harmon. Anal.,vol. 26, no. 3, pp. 301–321, 2008.

[21] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," IEEE Trans. Inf. Theory, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.