Miniaturization of Time-Gated Raman Spectrometer with a Concave Grating and a CMOS Single Photon Avalanche Diode

MINIATURIZATION OF TIME-GATED RAMAN SPECTROMETER WITH A CONCAVE GRATING AND A CMOS SINGLE PHOTON AVALANCHE DIODE

By

Zhiyun Li, M.Sc., B.Sc.,

A THESIS

SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

McMaster University

Hamilton, Ontario, Canada

DOCTOR OF PHILOSOPHY (2015) McMaster University

(School of Biomedical Engineering) Hamilton, Ontario

TITLE:	Miniaturization of Time-Gated Raman Spectrometer with a Concave Grating and a CMOS Single Photon Avalanche Diode
AUTHOR:	Zhiyun Li, M.Sc. (Xidian University)
SUPERVISOR:	Prof. M. Jamal Deen
NUMBER OF PAGES:	xvi, 152

Abstract

Raman spectroscopy possesses the important advantages of non-contact and non-destructive properties in chemical analysis applications. Commercial bench-top Raman spectrometers are bulky, expensive, and limited to laboratory use. Handset Raman spectrometers are commercially available, but still expensive. To extend its field applications such as water quality monitoring and pharmaceutical analyses, there is a growing need for cost reduction and system miniaturization of Raman spectrometers.

This work focuses on designing and building a compact and low-cost Raman spectrometer. A key issue of a Raman spectrometer is the detection of a weak Raman signal, especially when a strong fluorescence signal is present. Therefore, challenges in this work include system miniaturization and designing a low-cost and sensitive detection system to measure the weak Raman signal.

System miniaturization is approached using a concave grating based wavelength selector. A concave grating could perform both functions of light wavelength separation and focusing without the need for extra mirrors, reducing system size and complexity. The concave grating is designed and fabricated on plano-concave lenses by a low-cost custom holographic technique. Characterization of the custom concave grating proves its light wavelength separation and focusing properties.

Addressing cost reduction, a CMOS single photon avalanche diode (SPAD) is selected for its low cost and capability of detecting low intensity light. To achieve fluorescence suppression, the SPAD is operated in the time-gated (TG) mode. The TG-SPAD is implemented in a 130nm standard CMOS technology. Fast gating, generation and readout circuits are designed to gate the SPAD at frequencies up to 100MHz, with a short gate window of 3.5ns. Negligible afterpulsing probability (<1%) for hold-off times longer than 16ns is obtained.

A time-gated spectrometer prototype is built combining the concave grating and TG-SPAD. The system achieves timing resolution better than 60ps. Fluorescence suppression is observed by narrowing the detection window of the TG-SPAD, and Raman peaks of Rhodamine B are resolved by the system. The time-gated fluorescence lifetime measurement further proves the efficiency and fluorescence imaging capabilities of the proposed system.

Acknowledgments

Firstly, I would like to express my deepest appreciation to my supervisor, Prof. M. Jamal Deen, for giving me the opportunity to work on this interdisciplinary project. I am deeply grateful for his continuous guidance and support throughout the years of my Ph.D. study. I could never have reached the heights or explored the depths without his help and support. I feel very fortunate to work with such an outstanding professor.

I also would like to thank my committee members: Prof. P. Ravi Selvaganapathy and Prof. Qiyin Fang for their valuable advices and comments on my research. Their contributions to this research are greatly appreciated.

Many thanks to Dr. Ognian Marinov for keeping his door always open for technical supports and insightful discussions on carrying out experiments, data analysis, and paper reviewing. I am also grateful to Tianyi Guo, Darek Palubiak, Cong Feng, Hythm Afifi, Hani Alhemsi, Zeng Ceng, and Xiaoqing Zheng in the research team of the Nanoelectronics-Optoelectronics Research Lab, who gave me a wonderful experience over the years. Additionally, many thanks to Dr. Zhilin Peng for his help.

Last but not the least, I sincerely thank my family for their unceasing encouragement and support during my study. Thanks you for always being there for me.

Table of Contents

Abstract iii

Acknowledgments iv

Table of Contents v

List of Figures viii

List of Tables xiii

List of Symbols and Acronyms xiv

Chapter 1 1

Introduction and Applications 1

1.1 Applications 1

1.1.1 Water Quality Monitoring 1

1.1.2 Pharmaceutical Analysis 4

1.2 Existing Optical Technologies 5

1.2.1 Mid-Infrared Spectroscopy 6

1.2.2 Near Infrared Spectroscopy 7

1.2.3 Raman Spectroscopy 7

1.3 Challenges and Motivation 9

1.4 Time-Resolved Applications 11

1.5 Thesis Organization 13

1.6 Contributions 14

Chapter 2 16

Raman Spectroscopy and System Design 16

2.1 Raman spectroscopy 16

2.1.1 Scattering Theory 16

2.1.2 Raman Scattering Intensity 18

2.2 Instrumentation for Raman Spectroscopy 18

2.2.1 Excitation source 19

2.2.2 Wavelength Selector 21

2.2.3 Detector 24

2.3 Advanced Raman Techniques 30

2.3.1 Surface Enhanced Raman Spectroscopy (SERS) 31

2.3.2 Time-Gated Raman Spectroscopy 33

2.3.3 State-of the-Art Portable Raman Spectrometers and Design Challenges 36

2.4 System Design 38

2.4.1 Main System Components 38

2.4.2 System Operation and Specifications 40

2.5 Summary 42
Chapter 3 43
Wavelength Selector 43
3.1 Theoretical Background of Diffraction Grating 43
3.1.1 Dispersion 45
3.1.2 Free Spectral Range 45
3.1.3 Diffraction Limit 46
3.1.4 Diffraction Efficiency 47
3.2 Concave Grating Design 48
3.2.1 Diffraction Efficiency 48
3.2.2 Spectral Resolution 50
3.2.3 Flat-field Concave Grating Design 61
3.3 Concave Grating Fabrication and Characterizations 65
3.3.1 Grating Fabrication Techniques 65
3.3.2 Concave Grating Fabrication 68
3.3.3 Concave Grating Measurements and Comparison to Theory 71
3.3.4 Future Improvements 75
3.4 Summary 76
Chapter 4 77
Time-Gated Single Photon Avalanche Photodiode (TG-SPAD) 77
4.1 SPADs-Review and Theory 77
4.1.1 Free Running Operation 80
4.1.2 Time-Gated Operation 84
4.2 Design of TG-SPAD Pixel Circuit 86
4.2.1 SPAD Design and Characterization 86
4.2.2 On-chip Pulse Generator Design and Characterization 90
4.2.3 Design of TG-SPAD Pixel Circuit 92
4.3 Measurements of TG-SPAD Pixel Circuit 98
4.3.1 Functional Test of the TG-SPAD Pixel Circuit 99
4.3.2 Performance Tests of TG-SPAD Pixel Circuit 102
4.4 Summary and Future Improvements 110
Chapter 5 111
Time-Gated Spectrometer Implementation and Applications 111
5.1 Time-Gated Spectrometer Implementation 111
5.1.1 System Configuration 111
5.1.2 System Synchronization 114
5.1.3 System Timing Resolution 117
5.2 Raman Spectrometer Verification 119
5.2.1 Raman Spectrum Measurements 119
5.2.2 Analysis of the Time-Gated Raman Spectrum Measurement and Discussions 122
5.3 Fluorescence Lifetime Measurements 125

5.3.1 Fluorescence Decay Measurements 125
5.3.2 Analysis of Fluorescence Decay Measurement and Discussions 126
5.4 Future Improvements 131
5.5 Summary 132
Chapter 6 133
Conclusions and Recommendations for Future Work 133
6.1 Conclusions 133
6.2 Future Work 136
Appendix I: Design of Flat-Field Concave Gratings 139
Appendix II: Challenges and Solutions for the Fabrication of Concave Gratings 141
References 145

List of Figures

Figure 1.1: Schematic diagram of a flow cytometer 3

Figure 1.2: Principle of gas chromatography 4

Figure 1.3: Molecular structure [15] and Raman spectra of aspirin, commercial sample of aspirin, and aspirin inside plastic wrapper (30s acquisition time) [16] 6

Figure 1.4: Rayleigh and Raman scattering 8

Figure 1.5: Principle of time-gated fluorescence lifetime measurement 12

Figure 1.6: Scheme showing the "banana shaped" path of Near Infrared photons through the brain. The brain image without the source and detectors, is from [30] 12

Figure 2.1: Energy level diagram related to IR absorption, Raman scattering and fluorescence emission 17

Figure 2.2: System with (a) 90°; (b) 180° configurations for collection of the Raman scattering 19

Figure 2.3: Schematic of a monochromator 21

Figure 2.4: Simplified block diagram of the FT-Raman spectrometer 23

Figure 2.5: Basic structure of a CCD 26

Figure 2.6: Photon absorption efficiency with different wavelength: (a) Si; (b) Ge. (Xn: distance from surface to depletion region; W: depletion region width) 28

Figure 2.7: Cross section of a CMOS SPAD 30

Figure 2.8: Schematic illustration of SERS 32

Figure 2.9: Temporal variations of excitation, Raman scattering and fluorescence emission 34

Figure 2.10: Schematic of a Kerr gate system 35

Figure 2.11: Optical diagram of a concave grating based system 38

Figure 2.12: Time-gated operation of SPAD 40

Figure 2.13: Optical diagram of the proposed TG Raman spectrometer 41

Figure 3.1: Diffraction of a reflective grating 44

Figure 3.2: Spectrum of a white source dispersed by a diffraction grating 44

Figure 3.3: Diagram demonstrating the spectral overlapping of a grating 46

Figure 3.4: Simulation of the diffraction efficiency of concave gratings with different grating periods, at other conditions: 30deg incident angle of light, Au coated, 25mm radius 48

Figure 3.5: Simulation of the diffraction efficiency of a concave grating for different incident angles, at other conditions: 700nm grating period, Au coated, 25mm radius 49

Figure 3.6: Simulation of the wavelength response of gratings with surfaces coated by different materials (700nm grating period, 25mm radius, 30deg incident angle) 50

Figure 3.7: Rowland configuration of the concave grating system 51

Figure 3.8: Geometry of a concave grating: A-point source, O-grating center, B-image of point source in image plane, P-arbitrary point on grating surface 53

Figure 3.9: Spot diagrams of a point source with different aberration (λ=600, 602nm): (a) astigmatism; (b) coma; (c) spherical aberration; (d) image with all aberrations 55

Figure 3.10: Spot diagram of: (a) Illuminated area of a point source on the concave grating surface; (b) Image of the corresponding point source 57

Figure 3.11: Situation 1 (25mm radius, 700nm grating constant, 0.12 NA): (a) Power series coefficients with incident angle. (b) Aberration related spectral resolution when considering different Taylor series coefficients. 58

Figure 3.12: Situation 2 (25mm radius, 700nm grating constant, 0.12 NA): (a) Power series coefficients with incident angle. (b) Aberration related spectral resolution when considering different power series coefficients. 59

Figure 3.13: Total spectral resolutions when (a) the height h is close to the width w; and (b) the height h is much smaller than the width w. 60

Figure 3.14: Horizontal focal curve of constant space concave grating (Incident 35deg, grating constant 700nm, radius 25mm, wavelength 550-650nm): Rowland circle (Red), grating (blue), focal curve (green) 61

Figure 3.15: Dependence of H_{20} on the incident angle α and reference wavelength λ_0 63

Figure 3.16: Sum of the first three high order terms in Eq. (3.26) and the modified linear focal curve (Red) corresponding to the original circular focal curve (blue) with for wavelength band from 550nm to 650nm (Grating constant 700nm, incident angle 35deg): (a) Reference wavelength=600nm; (b) Reference wavelength=800nm 64

Figure 3.17: Optical diagram of concave micro-grating spectrometer system (Reference wavelength=900nm) 65

Figure 3.18: Schematic of holographic grating fabrication 66

Figure 3.19: Replication of a master grating. The grating structure (e) is from [119]. 67

Figure 3.20: Optical setup of the holographic fabrication: (a) Common setup; (b) Setup in this design 68

Figure 3.21: Glass substrate from Thorlabs: (a) Plano-concave lens (LC1258); (b) Plano-convex lens (LA1257); (c) Grating on Plano-concave lens; (d)Grating on Plano-convex lens 70

Figure 3.22: SEM image of the concave grating surface 71

Figure 3.23: Optical diagram of the Rowland configuration with source, grating and detector....72
Figure 3.24: Experimental setup of the concave grating test 72
Figure 3.25: Dispersion of the concave grating of a white light 73
Figure 3.26: (a) Schematic of the focusing property measurement setup; (b) Images taken off and at the Rowland circle; (c) Intensity distribution of images 73
Figure 3.27: (a) 3D model designed by Autodesk inventor; (b) Optical setup with the 3D model74
Figure 3.28: Spectrum resolved by the concave grating 75
Figure 4.1: Current-voltage (IV) characteristic of avalanche photodiode operated in Geiger mode 78
Figure 4.2: (a) SPAD with passive quenching and recharge circuit; (b) Equivalent circuit of passive quenching; (c) Equivalent passive reset circuit 81
Figure 4.3: Simulation of the SPAD cathode potential $V_o(t)$ with passive quenching and recharge. 82
Figure 4.4: [131] (a) Basic diagram of a mixed passive-active quenching circuit; (b) Cathode voltage waveform 83
Figure 4.5: (a) Basic diagram of SPAD with time gated control circuit; Timing diagram of the gating signal (b) and the output of the AC pick-up circuit (c) 85
Figure 4.6 Block diagram of the proposed TG-SPAD pixel circuit 86
Figure 4.7 Cross section view of photodiode implemented in CMOS technology 87
Figure 4.8: Layout top view of the N+/P-well diode (a) and P-well/Deep N-well diode (b) 87
Figure 4.9: (a) I-V measurements of the two diodes with P region grounded; (b) Extraction of the dynamic resistance of the N+/P-well diode 88
Figure 4.10: Optical response measurements: (a) N+/P-well diode; (b) P-well/Deep N-well diode 89
Figure 4.11: (a) Circuit topology of the on-chip pulse generator; (b) Waveform in the pulse generator by Cadence 90
Figure 4.12: (a) Pulses measured from six chips; (b) Pulse measured with different temperatures 92
Figure 4.13: Schematics of the (a) proposed TG-SPAD pixel circuit, (b) on-chip gating control circuits 93
Figure 4.14: Timing diagram of the TG-SPAD pixel circuit (Simulated by Cadence Virtuoso) ..94
Figure 4.15: (a) Schematic diagram and (b) timing diagram of the wave shaping circuit 97
Figure 4.16: Layout view of the designed chip with the prototype of the time-gated SPAD front-end 98

Figure 4.17: (a) Micrograph of the fabricated TG-SPAD and the packaged chip; (b) PCB with DC and RF connectors and by-pass capacitors 99

Figure 4.18: Outputs of the TG-SPAD pixel circuit: (a) Simulation; (b) Measurement 100

Figure 4.19: Output of TG-SPAD pixel circuit with in-pixel wave shaping circuit 101

Figure 4.20: Histogram of the output pulse width: (a) pixel without wave shaping; (b) pixel with wave shaping 101

Figure 4.21: (a) Temperature and excess bias dependence of dark count probability per gate window (DCP_{GW}) of the TG-SPAD front-end; (b) Arrhenius plot of DCP_{GW} 103

Figure 4.22: Schematic (a) and equivalent circuit (b) of the TG-SPAD pixel circuit 106

Figure 4.23: Measured dark count probability per gate with deadtime 107

Figure 4.24: Measured afterpulsing probability of the TG-SPAD (line with small circles), and comparison with reported data (symbols only) from other publications (Key to references: A [150], B [151], C [154], D [155], E [156], F [157], and G [148]) 108

Figure 4.25: Schematic diagram of the PDE measurement 109

Figure 4.26: Measurement of PDE as a function of excess bias 109

Figure 5.1: Optical setup of the system in a Rowland configuration 113

Figure 5.2: (a) Schematic diagram of the time jitter measurement; (b) Histogram of the time between laser trigger signal and laser emission 115

Figure 5.3: (a) Block diagram; (b) experimental setup of the optically synchronized time-gated spectrometer 116

Figure 5.4: Photon arrival time measurement of a 7ps pulsed laser 118

Figure 5.5: Light attenuation in Raman spectrometer 119

Figure 5.6: (a) Illumination of sample carried in plastic cuvette; (b) Spectrum measured by a commercial spectrometer (OEM-400, Newport, Irvine, CA, 0.3nm spectral resolution) 121

Figure 5.7: Timing diagram of the measurements 122

Figure 5.8: Spectrum of Rhodamine B measured with different detection window 123

Figure 5.9: Fluorescence background (a) and Raman peaks (b) measured at different detection windows 124

Figure 5.10: Principle of the time-gated fluorescence lifetime measurement 126

Figure 5.11: Fluorescence decays measurements of Rhodamine 6G and Rhodamine B 127

Figure 5.12: Photon arrival time distributions within each gate window of different time delays 128

Figure 5.13: Histogram and cumulative probability of photons measured with a delay time of 4ns after the laser pulse excitation-Rhodamine 6G 129

Figure 5.14: Fluorescence lifetime extracted with different cumulative probabilities of photons (Rhodamine 6G) 130

Figure II.1: AFM image of a grating on a silicon substrate: (a) 2D surface image; (b) 3D image 142

Figure II.2: (a) Double concave watch glass sample; (b) Glass sample after fabrication 143

Figure II.3: 3D AFM image of the grating on concave glass substrate (measured at the top of the substrate) 144

List of Tables

Table 1.1 Water applications and associated quantities for evaluation of the water quality [6].....2

Table 1.2 Comparison of features, instrumentations, and applications of different types of vibrational spectroscopy.....9

Table 2.1: Commercial lasers used in Raman spectrometers and applications.....20

Table 2.2: Miniaturized spectrometers with nanometer spectrum resolution 23

Table 2.3: SERS in water contaminants application 33

Table 2.4 Summary of characteristics of recently developed time-gated Raman systems 36

Table 2.5. Commercially developed portable Raman spectrometers.....37

Table 2.6 Target specifications of proposed system 42

Table 3.1: System specifications of the spectral resolution measurement.....75

Table 4.1: Parameters used for simulation of the passive quenching and recharge circuit.....82

Table 4.2. Comparison of activation energies and other performances of different SPADs 104

Table 5.1: Specifications of Passat Compiler 355 pulsed laser.....112

Table 5.2: Comparison of performance characteristics and applications of state-of-the-art CMOS TG-SPADs 118

Table II.1 Grating fabrication parameters 141

List of Symbols and Acronyms

Symbols

I_R	Intensity of Raman signal
I_0	Intensity of excitation source
P	Dipole moment
α	Incident angle
v_0	Frequency of excitation
d	Grating constant
G	Groove density
λ	Wavelength
β	Diffraction angle
m	Diffraction order
S	Entrance slit width
$\Delta\lambda_{Entrance}$	Spectral resolution caused by entrance slit
$\Delta\lambda_{Diffraction}$	Spectral resolution caused by diffraction limit
$\Delta\lambda_{Aberration}$	Spectral resolution caused by aberration
V_{BR}	Avalanche breakdown voltage
V_{ex}	Excess bias
R_Q	Quench resistor
R_D	Space charge region resistance
C_D	Space charge region capacitance
W	Space charge region width
t_{win}	Gate window
t_{RST}	Reset time
t_{off}	Hold-off time
DCP_{GW}	Dark count probability per gate window

T	Absolute temperature
k	Boltzmann's constant
E_g	Energy bandgap
U_{SRH}	Shockley-Read-Hall generation rate
q	Electron charge

Acronyms

AC	Alternating current
AP	Afterpulsing probability
BOD	Biochemical oxygen demand
CARS	Coherent Anti-Stokes Raman Spectroscopy
CCD	Charge-coupled device
CMOS	Complementary metal oxide semiconductor
COD	Chemical oxygen demand
DC	Direct current
DCR	Dark count rate
DSM	Deep sub-micron
FLIM	Fluorescence lifetime imaging
FF	Fill factor
FON	Film over nanosphere
FTIR	Fourier transform infrared spectroscopy
FWHM	Full-width half-maximum
GC	Gas chromatography
GR	Generation-recombination
ICCD	Intensified charge-coupled device
IV	Current-voltage
LIDAR	Light detection and ranging
LOD	Limit of detection
MIR	Middle Infra-red

NA	Numerical aperture
NFET	N-type field-effect transistor
NRS	Normal Raman spectroscopy
OPD	Optical path difference
SERS	Surface enhanced Raman spectroscopy
PDE	Photon detection efficiency
PFET	P-type field-effect transistor
PMT	Photo multiplier tube
POP	Persistent organic pollutants
QE	Quantum efficiency
SRH	Shockley-Read-Hall
SPAD	Single photon avalanche diode
TCSPC	Time-correlated single photon counting
TOC	Total organic carbon
UV	Ultraviolet

Chapter 1

Introduction and Applications

Optical characterization techniques were, and are still being actively researched, because they provide valuable information by non-destructive and non-contact methods of testing. In particular, vibrational spectroscopy probes the fundamental vibrations of molecules, allowing for identification of molecular structures. Therefore, optical characterization can be used for a wide range of applications, including material science, food safety, environmental and pharmaceutical analyses. However, due to the bulky and expensive instruments such as a Raman spectrometer, most optical characterizations are carried out in laboratories. Portable instruments are commercially available for field applications of vibrational spectroscopy, but are still expensive. In this situation, system miniaturization and cost reduction of the instruments are indispensable to extend optical characterization techniques to the field applications. Several examples requiring field tests are given next, followed by a comparison of commonly used optical characterization techniques.

1.1 Applications

1.1.1 Water Quality Monitoring

Water is more crucial than any other resources on Earth for human. According to [1], among the total volume of water on Earth, only 2.5 percent or about 35 million km^3 is fresh water. In addition, most fresh water is not usable. They are in the form of permanent ice or snow, existing in areas such as Antarctica and Greenland. Major sources of human usable water are lakes, rivers, soil moisture and shallow groundwater basins. The human usable water is less than a percent of the freshwater on Earth. Therefore, there is a water crisis worldwide [1].

According to the Global Water Partnership [2]: the global water usage has doubled from 1960 to 2000. However, about a third of the world's population lacks of sufficient access to safe drinking water and sanitation to meet their basic needs (Pacific Institute 2007), 900 million people rely on unimproved/untreated drinking water supplies (WHO/UN-water, 2008). Today, there is a water crisis not only from insufficient water resources to satisfy our needs, but also from the insufficient management of water resources [3]. To improve water resources management, timely monitoring of water quality is of great importance.

Water quality indicates the suitability of a water resource to meet specific needs. For example, drinking water should have very low or zero concentration of chemicals and micro-organisms harmful to health. Water for irrigation in agriculture should have low sodium-content [4]. To monitor water quality, various quantities are measured, including chemical concentration of microbiological, organic and inorganic contaminants, physical and other indicators, such as pH, temperature or turbidity [5]. Table 1.1 outlines the suites of substances that are monitored for water for different uses [6].

Table 1.1 Water applications and associated quantities for evaluation of the water quality [6]

Purpose of use	Human Health Drinking water	Agriculture	Municipal/ Industrial	Ecosystem Stability	Tourism & Recreation
Parameters	Total Coliform Faecal Coliform Pathogens POPs Turbidity Trace metals	Nutrients Nitrogen Phosphorus Salinity Chlorophyll A Pathogens	BOD COD Heavy Metals (particularly in Sediment)	Temperature pH - acidity Conductivity Major ions Oxygen Suspended Solids Biodiversity	Parasites Pathogens

BOD: Biochemical Oxygen Demand (quantity of oxygen used by microorganism in the oxidation of organic matter)
COD: Chemical oxygen demand (quantity of oxygen used in the oxidation of organic matter and inorganic chemicals)
POPs: Persistent organic pollutants (toxic chemicals persistent in the environment, adversely affect human health)

Water quality can be tested either on site or in laboratory. With precise control of test conditions and advanced instruments, laboratory tests can provide a low limit of detection (LOD). Advanced analytical instruments are commercially available for precise evaluation of different water quality indicators. For example, analyzers of TOC (total organic carbon), gas chromatography and enzyme-based systems with immunoassays are able to detect specific chemical or biological contaminants, and flow cytometers have been used for detection of different microbial contaminants [7], [8].

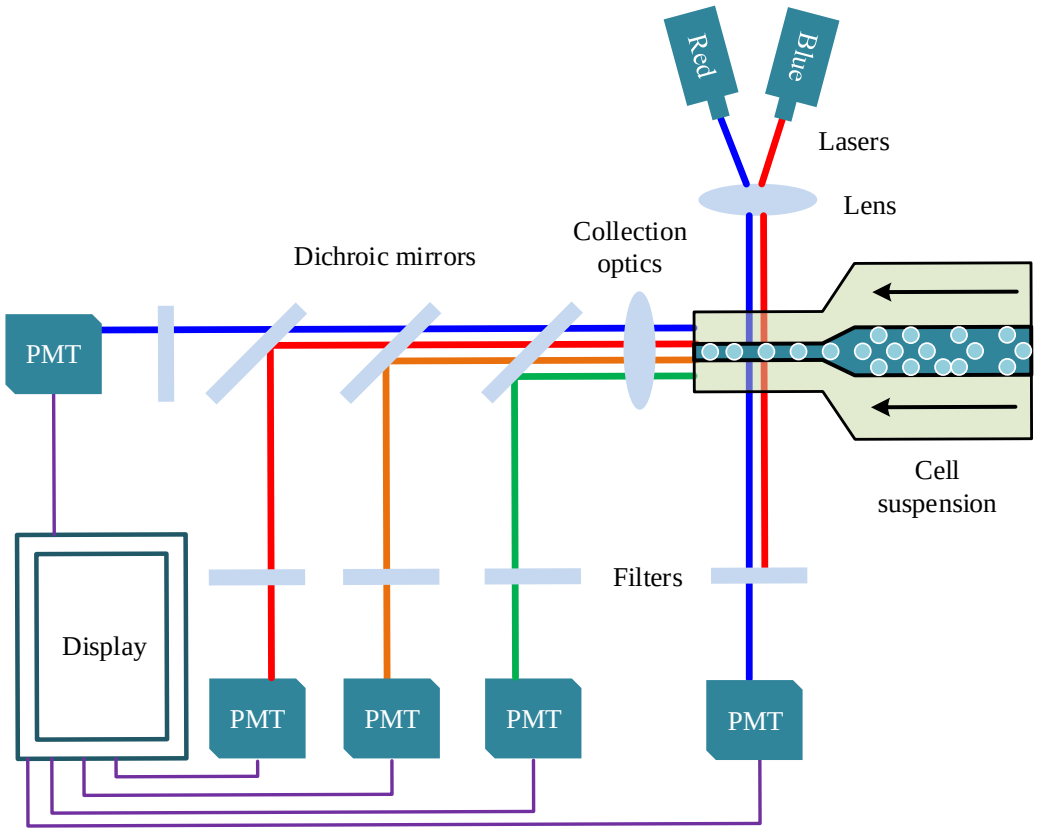


Figure 1.1: Schematic diagram of a flow cytometer

Figure 1.1 shows a functional diagram of a flow cytometer. The cell suspension passes through a flow chamber. Meanwhile, a laser beam is focused onto the flowing cells by a lens. When the laser light strikes a cell, the light is forward and side scattered at the same wavelength as the incident laser source. Photomultiplier Tubes (PMTs) are used to detect the scattered signals and count the number of cells that passed through the flow chamber. For a sample volume less than a millilitre, up to 1000 particles per second can be counted [9], [10]. In addition, cells can be stained with fluorescence dyes, so that fluorescence signals can also be detected at wavelengths different from the incident wavelength. As shown in Figure 1.1, different fluorescence dyes can be attached to different types of cells, and a set of dichroic mirrors is used to separate different fluorescence signals according to their wavelengths. The utilization of fluorescence dyes is useful to distinguish different types of cells.

Tests with advanced analytical instruments offer higher precision and better LOD. However, most advanced instruments are for laboratory use and require skilled personnel to conduct the tests. Moreover, extra sample treatment is required before measurement. Therefore, they are expensive

and require transportation of the sample to the laboratory. These factors result in a delay in the responses to contamination events, which is detrimental to water safety and public health.

In contrast, field tests are carried out in situations in which an immediate and fast monitoring of water quality is required. These tests can provide a rough estimation of water quality or used for water quality screening. Also, since water samples are immediately analyzed, then field test avoids sample contamination or degradation during sample transportation or storage. However, due to the poor LOD, field tests are mainly for simple measurements, e.g., temperature, pH or conductivity. Thus, there is a growing and urgent need for a robust, low-cost, continuous, fast and accurate field detection system for water quality monitoring.

1.1.2 Pharmaceutical Analysis

Pharmaceutical analysis is defined as the process of identification and quantification of the chemical compositions and impurity contents during the formulation of pharmaceutical products, through investigation of bulk drug materials, intermediates, drug products, drug formulation, and degradation products. The primary goal of pharmaceutical analysis is to ensure the quality of pharmaceutical products. Well-planned testing with suitable methodology and instrumentation can help build quality into a pharmaceutical product [11].

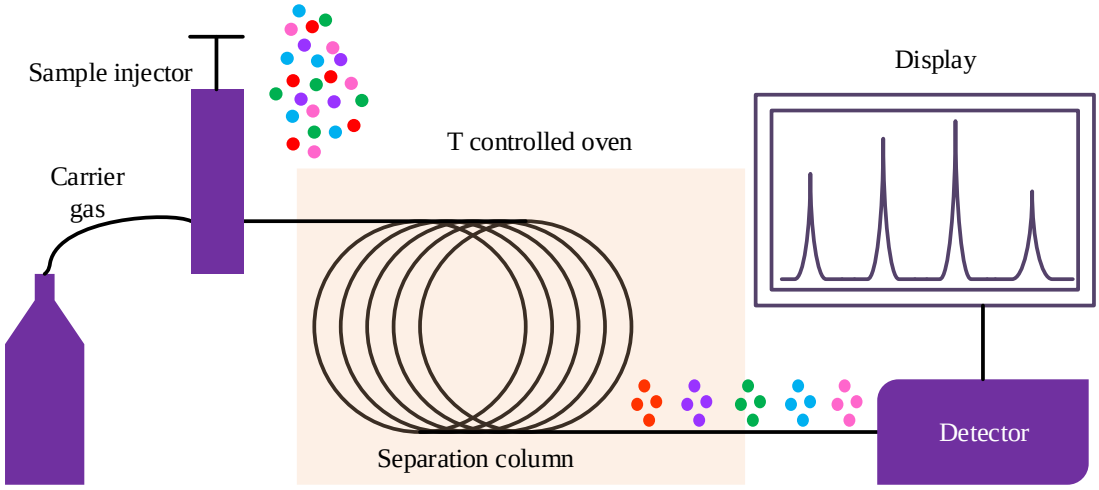


Figure 1.2: Principle of gas chromatography

To assess the quality of drug products, different analytical techniques such as chromatographic or spectroscopic techniques have been applied for pharmaceutical analysis. Figure 1.2 shows the principle of gas chromatography (GC). Gas chromatography is an analytical technique for separation and analysis of evaporated chemical compounds. The separation of chemical

compounds in GC is achieved by using a mobile and a stationary phase. The mobile phase is comprised of an inert gas, such as nitrogen, argon, or helium. The stationary phase consists of packed columns with large surface area, on which the liquid stationary phase is coated, and is installed in a temperature controlled oven. During the measurement, the gas or liquid sample is injected from the sample injector, vaporized and carried by the mobile carrier gas. When traveling through the stationary phase, chemicals of different constitutes are separated along the long separation column because of their different retention times, depending on their solubility in the stationary phase at the given temperature. A detector is installed by the end of the stationary phase to record the arrival time and the quantity of each component [12].

Comparing with other analytical techniques, spectroscopic techniques such as vibrational spectroscopy benefit the pharmaceutical analysis from the following aspects:

1) *in situ* spectra acquisition: non-contact measurement, with a little or no sample preparation;
2) irrespective of sample state: measurements can be carried out of any sample state, gas, liquid, solution, or solid;
3) small sample amount: an important property since many drugs are formulated as microcrystals in early stage;
4) tests conducted with optical fibers: measurements can be carried out in a dangerous environment or outdoors.

Figure 1.3 shows the Raman spectra of aspirin measured in 3 different conditions with the same Raman spectrometer and at fixed acquisition time. From the results we can see that the aspirin tablets with and without the plastic wrapper give identical Raman spectra, which proves the non-contact *in situ* measurement property of Raman spectroscopy.

1.2 Existing Optical Technologies

In the past decades, high speed detection systems for on-site water quality monitoring and pharmaceutical analysis have attracted much interest from industry and academia [7], [13]. Vibrational spectroscopy has been extensively studied because of its non-contact and non-invasive properties, with minimum sample preparation, and for rapid detection and identification of different chemical and microbial samples [5], [14], as the example shown in Figure 1.3.

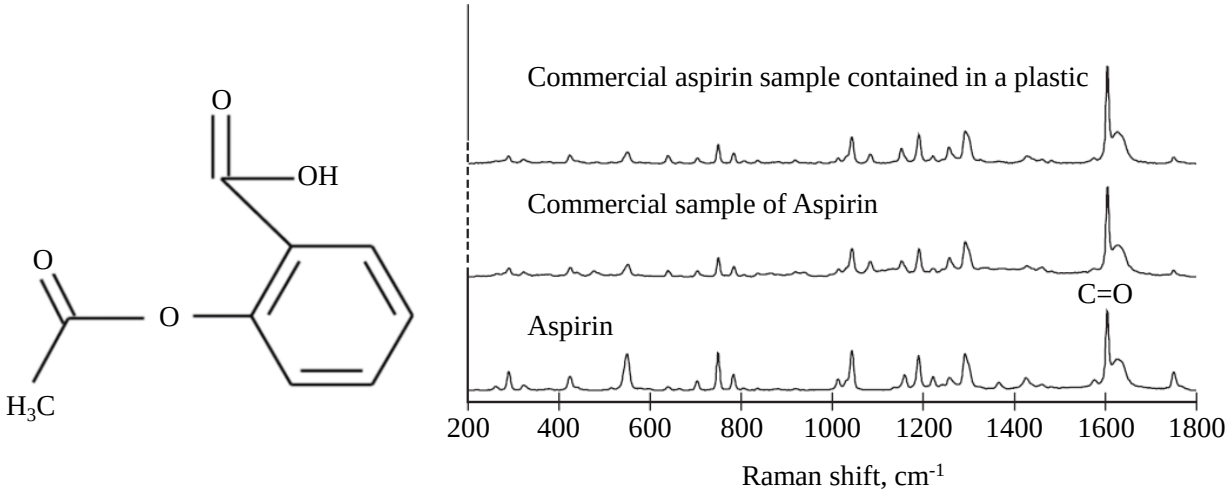


Figure 1.3: Molecular structure [15] and Raman spectra of aspirin, commercial sample of aspirin, and aspirin inside plastic wrapper (30s acquisition time) [16]

Vibrational spectroscopy refers to the measurement of the vibrational energy levels of the chemical bonds in a molecule. The vibrational spectrum is rich in information regarding the chemical composition of the sample, so it has been applied for chemical analysis such as water quality and material science. Infrared (IR) and Raman Spectroscopies are the two most commonly used vibrational spectroscopy techniques for chemical and biological analysis. Within IR spectroscopy, both near infra-red (NIR) spectroscopy and middle infra-red (MIR) spectroscopy are used in chemical analyses.

1.2.1 Mid-Infrared Spectroscopy

Fundamental transitions (of electrons) in molecules usually occur by absorbing photons in mid-infrared (MIR) wavelength range. MIR (2.5-16μm) spectroscopy is a well-established vibrational spectroscopy for analysis of molecular structure. MIR spectroscopy measures either the reflection or transmission of optical signals of a target, evaluating the photon absorption in the target. Fourier transform infrared spectroscopy (FTIR) is widely used to measure MIR spectra. In a FTIR spectrometer, a MIR source is used to illuminate the sample. The reflected or transmitted signal is collected by light collection optics, and is directed to an interferometer. At the output of the interferometer, the optical signal is detected by an IR detector, from which an interferogram is obtained. Finally, the MIR spectrum is obtained by the Fourier transform of the interferogram.

Water is a strong MIR absorber, so special care is usually taken when measuring liquid samples by FTIR. To reduce the IR photon absorption by water, the travel length of IR photons in a water sample is reduced. For instance, the thickness of sample holder is designed in the range of

micrometers to ensure that a signal of sufficient intensity is obtained. Alternatively, the attenuated total reflection (ATR) method is used to strengthen the interaction between the IR signal and sample [17]. Finally, some pre-concentration techniques have been used to enhance the absorption in the sample [18].

1.2.2 Near Infrared Spectroscopy

Near infrared (NIR) spectroscopy covers a wavelength range from 780nm to 2500nm. Since water absorption in the NIR region is weaker than in the MIR region, then NIR spectroscopy is more suitable for direct measurement of liquid samples. A NIR spectrometer contains a light source for sample illumination, a spectrograph to resolve the reflectance or transmittance spectra, and a detector to measure the spectra. The light source usually is a lamp providing broadband excitation in the NIR region. For the spectrograph, both dispersive and non-dispersive units are used in commercial NIR spectrometers. For detector, silicon-based charge-coupled devices (CCDs) are used to acquire spectra at wavelengths below 1000nm. However, for wavelength > 1000nm, indium gallium arsenide (InGaAs) photodetectors are commonly used because at these wavelengths, silicon is transparent.

In contrast to MIR spectroscopy, NIR spectroscopy originates from the overtones and combinations of fundamental vibrations [19]. The NIR absorption is typically in broad and overlapping bands, and the intensities of the bands are weaker than that of the fundamental absorption bands in MIR. Because of the combinations of vibrational modes, NIR spectra are complex and it is often difficult to directly link a broad absorption band to a particular chemical bond [5]. Although NIR spectroscopy is not as well-established as MIR spectroscopy, NIR spectroscopy is less expensive and increasingly considered as a promising technique for characterization of inorganic materials [20], [21].

1.2.3 Raman Spectroscopy

Raman spectroscopy relies on the inelastic scattering of a monochromatic light from a molecule. Since water is a weak Raman scatterer, Raman spectroscopy is superior to other vibrational spectroscopies, particularly for measurements of liquid samples in pharmaceutical analysis, biomedical diagnosis and tissue imaging [22]-[24]. The instrumentation for Raman spectroscopy is similar to that for NIR spectroscopy, except for the light source. Unlike in MIR and NIR spectrometers, a monochromatic source, usually a laser, is used in a Raman spectrometer. The

wavelength of the excitation lasers ranges from ultraviolet (UV) to visible, or even NIR regions. Since the Raman spectra contain information about fundamental vibration modes of a molecule, Raman spectroscopy is said to be complementary to MIR spectroscopy. Figure 1.4 shows the schematic of Rayleigh and Raman scattering. Details of Raman scattering is given in chapter 2.

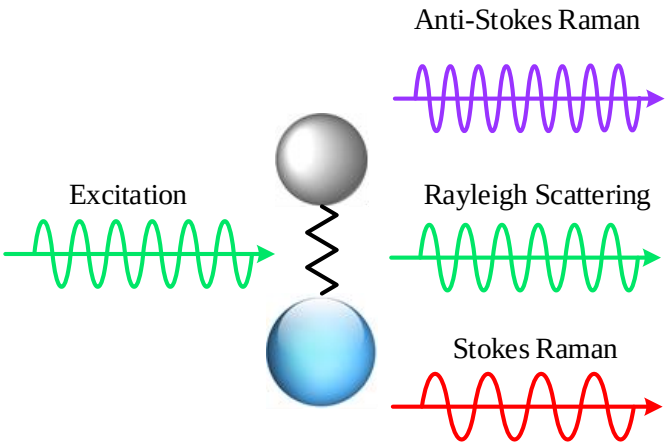


Figure 1.4: Rayleigh and Raman scattering

Although Raman scattering was observed as early as in 1928, Raman spectroscopy was slowly developed before the 1970s due to several reasons. First, Raman scattering is a very weak process. Sensitive instrumentations are required to detect the weak Raman signal. However, these instruments were not available in the early days. Second, when a molecule is excited by a photon with high enough energy (UV or visible), in addition to the Raman signal, a fluorescence signal whose intensity is much stronger than the intensity of the Raman signal may also be emitted. With the selected wavelength of excitation source, if the fluorescence emission band overlaps with the Raman peaks, the fluorescence signal can easily overwhelm the Raman signal, making it very difficult to resolve the Raman signal.

Following the development of laser sources, Raman spectroscopy was first used to detect organic contaminants in water in 1970 [25]. However, because of the low Raman scattering efficiency, together with the strong accompanying fluorescence and the lack of sensitive instruments, Raman spectroscopy was not widely employed before the 1980s. With the advances in silicon-based detector arrays of high detection efficiency, low noise, and the increasing availability of stable and high power laser diodes, Raman spectroscopy has progressed quickly since the late 1980s [26]. In addition, with the advent of advanced Raman techniques such as the surface enhanced Raman spectroscopy, the limit of detection has been significantly improved, making Raman spectroscopy suitable for measuring chemical contaminants in water.

Overall, vibrational spectroscopy is a non-contact, non-destructive technique for chemical and biological analysis. It has been successfully applied to different fields, e.g., in the pharmaceutical industry and for water quality monitoring. Table 1.2 compares the features, instrumentation, and applications of the three types of vibrational spectroscopy.

Table 1.2 Comparison of features, instrumentations, and applications of different types of vibrational spectroscopy

Technique	Features	Instrumentation	Applications
MIR Spectroscopy	• Absorption spectroscopy • Fundamental vibration mode • Restriction in liquid sample	• Polychromatic source, MIR • Interferometer, filter • IR detector	Pharmaceutical and agricultural applications, food science, microbial cells
NIR Spectroscopy	• Absorption spectroscopy • Overtone and combination	• Polychromatic source, NIR • Interferometer, grating • CCD, PMT	Clinical chemistry, near infrared tomography, industrial process control, water quality
Raman Spectroscopy	• Scattering spectroscopy • Fundamental vibration mode • Low intensity	• Monochromatic source, UV, visible, NIR • Grating , interferometer • CCD, PMT	Pharmaceuticals and cosmetics, geology and mineralogy, semiconductor materials characterization

CCD: charge-coupled device
PMT: photo-multiplier tube
APD: avalanche photodiode
SPAD: single photon avalanche diode

1.3 Challenges and Motivation

Although vibrational spectroscopy is a powerful technique for chemical analysis of water quality and pharmaceutical samples, there are several limitations for field applications. MIR spectroscopy is the most developed vibrational spectroscopy. Since water is a strong MIR absorber, even though water absorption of MIR photons can be weakened by using specially designed sample holders, they also reduce the quantity of samples to be measured. In addition, commercial bench-top FTIR spectrometers are bulky, expensive, and limited to laboratory use. Portable FTIR spectrometers are commercially available (Thermo Scientific- TruDefender FT), but they are not suitable for liquid samples.

Compared with MIR spectroscopy, NIR spectroscopy suffers less from water absorption, and simpler sample preparation is required. Measurements can be carried out by fiber optic probes, so in principle, NIR spectroscopy is suitable for field applications. However, because NIR spectroscopy originates from overtones and combinations of fundamental vibrational modes. NIR spectra do not reflect directly the chemical composition information of the measured sample. To obtain useful information from the NIR spectra, complex statistical processing of the NIR spectra is needed, which requires input from experienced experts. Statistical techniques for NIR spectra

analyses are still under development in both industry and academia. Consequently, NIR spectroscopy is inapplicable for field applications at present.

Among the vibrational spectroscopic techniques, Raman spectroscopy is the most promising technique for measurement of liquid samples. Since the sample preparation (if any) is relatively simple, then Raman spectroscopy can be arranged for *in situ* applications. Moreover, Raman spectra contain information for the fundamental molecular vibrational modes, and this information can be understood without complicated data analysis. The main issue hampering the field application of Raman spectroscopy is the expensive instrumentation. A commercial bench-top Raman spectrometer usually costs more than $100,000. Portable Raman spectrometers are commercially available, but these handheld Raman spectrometers are also expensive (~$30,000), in particular those with fluorescence rejection capabilities. Therefore, to extend the field application of Raman spectroscopy, cost reduction of the Raman spectrometer is critical.

To be applied for field applications, three conditions have to be satisfied. First, the system must be capable of measuring chemical or biological samples, which is not a problem for Raman spectroscopy. Second, the system has to be compact and easy to use. Portable Raman spectrometers are commercially available, instilling confidence that the design and fabrication of a compact Raman spectrometer is possible. Third, the system should be inexpensive, which is one of the most challenging conditions for Raman spectrometers at present.

This research is devoted to the design and characterization of components for a low-cost Raman spectrometer. A Raman spectrometer consists of four basic units: an excitation source, optics for light illumination and collection, a wavelength selector, and a detection system. To measure the Raman spectrum of a sample, the wavelength of the laser source must be very stable and have relatively high power, the wavelength selector should offer high spectral resolution, and the detection system should be of high speed and very sensitive. These requirements contribute to the high cost of a Raman spectrometer.

Targeting size and cost reductions of the proposed system, this research focuses on the design and characterization of the wavelength selector and detector that are fabricated by inexpensive technologies. System miniaturization was addressed by using a wavelength selector with a concave grating. This grating was designed and fabricated by a custom low-cost holographic technology. Also, the concave grating-based wavelength selector is very compact, and provides good spectral resolution for Raman spectroscopy application.

A single photon avalanche diode (SPAD) was designed and used to detect the Raman signals. To suppress the fluorescence signal, the SPAD was operated in a fast time-gated mode, which allowed for temporal separation of the instantaneous Raman emission from the delayed and unwanted photons of the background fluorescence emission, when the sample is illuminated by a narrow pulse laser. Fast gating circuits were designed to drive the SPAD properly and a time-gated readout circuit was used to sense the detection events from the SPAD. The on-chip implementation of the time-gated SPAD in standard CMOS technology significantly reduces the cost of the system.

Combining the customized wavelength selector and the high-speed CMOS SPAD, a prototype time-gated spectrometer was developed, characterized, and the results are presented in this thesis and in some publications. The Raman peaks of a fluorescence dye were successfully resolved despite the presence of a strong fluorescence background.

1.4 Time-Resolved Applications

In addition to the spectroscopic application, and since a pulsed laser is used, the proposed time-gated system can also be used for time-resolved applications. Time-resolved measurements have been widely researched for their capability to provide temporal distribution of photons in applications such as diffuse optical tomography [27] and cellular imaging [28]. In a time-resolved measurement, fast photon detection is provided following the excitation of a sample by a narrow pulse laser. The time-correlated single photon counting (TCSPC) technique and time-gated photon detection are commonly used in time-resolved applications.

A typical time-resolved application of time-gated photon detection is fluorescence lifetime imaging (FLIM), which generates an image on basis of the differences in the excited state decay rates of a fluorescence sample. Since the fluorescence signal has an exponential decay, narrow gate windows can be used to count the number of photons emitted within the gate window, as shown in Figure 1.5. By counting photons in each gate window, a histogram can be built to reconstruct the curve of fluorescence emission and extract the fluorescence lifetime. FLIM has been applied to various fields of researches on tissues and cells. The fluorescence lifetime is independent of sample concentration and excitation source intensity, but it changes with the environment of the fluorophore, such as temperature and pH. Therefore, the fluorescence lifetime can be used for local environment sensing [29].

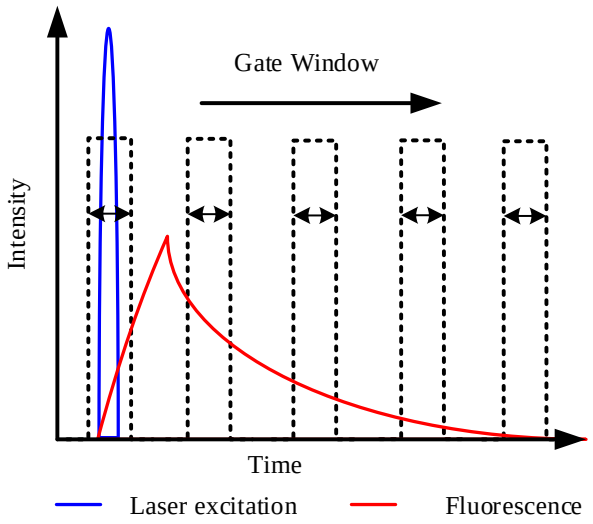


Figure 1.5: Principle of time-gated fluorescence lifetime measurement

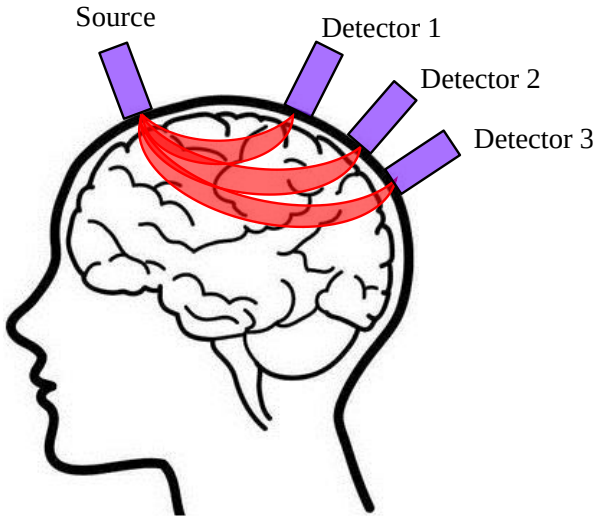


Figure 1.6: Scheme showing the "banana shaped" path of Near Infrared photons through the brain. The brain image without the source and detectors, is from [30]

In addition to FLIM, time-gated photon detection can also be used in time-resolved near infrared spectroscopy (NIRS) for functional brain imaging [31], [32]. In NIRS, brain activity is measured by optically probing variations in the concentrations of its oxygenated and deoxygenated haemoglobin. As shown in Figure 1.6, in NIRS, photons from a pulsed laser source are injected into the brain by a single fiber or a bundle of fibers. After penetrating inside the brain to a certain depth, the reflected photons are collected and detected at positions of predetermined source-detector separations. In time-resolved NIRS, the time-gated photon detection is used to measure the distribution of photon time of flight, from which the mean optical path length of photons can

be calculated. Moreover, "early photons" scattered by surface regions can be removed by time-gated detection. This is especially important when small source-detector separation is used [31].

1.5 Thesis Organization

This thesis is organized as follows. In chapter 1, an introduction to several state-of-the-art optical characterization techniques and their applications are provided. The motivation of focusing on optical characterization techniques as well as important challenges are discussed. Then, a brief description of the thesis and the main contributions of this research are presented.

In chapter 2, a review on Raman spectroscopy is presented. This review includes the theory, development of instrumentation, and several advanced techniques for Raman spectroscopy. After summarizing the limitations of current Raman instrumentations and challenges in the design of Raman systems, an overall system design is provided, addressing the major requirements for the system components and the key specifications of each component.

A key component in a Raman system is a wavelength selector. In chapter 3, theory and design of the wavelength selector is presented. The main component in a wavelength selector is the diffraction grating, so diffraction theory and related parameters are introduced. Then, the design considerations for a concave grating are given, followed by the fabrication details and results from performance characterization.

In chapter 4, a time-gated SPAD is designed for implementation in a 130nm standard CMOS technology. To achieve fast gating of the SPAD, a single avalanche diode was characterized first, and then an on-chip gating circuit was designed to provide a narrow gate window of 3.5ns for the SPAD. The chip was specially designed to extract photon arrival time information within each gate window. The functions of the chip were tested in dark and under illumination by a halogen lamp to obtain random photon arrival times, followed by a full characterization of the SPAD's performance.

Combining the concave grating and the time-gated CMOS SPAD, the setup and synchronization of the time-gated spectrometer prototype are discussed in chapter 5. The timing resolution of the entire system was measured using a narrow pulsed laser. To test the efficiency of the system for fluorescence rejection, Raman spectra of a fluorescence dye were measured. Fluorescence lifetimes of two fluorescence dyes were also measured by this time-gated system in

order to confirm the timing accuracy and other features of the time-gated SPAD for applications in time-resolved measurements.

In chapter 6, a summary of the research and recommendations for future improvements for system miniaturization and cost reduction are presented.

1.6 Contributions

This thesis aims at designing and building a compact and low-cost Raman spectrometer. Major contributions of this work are as follows:

1. Simplified algorithms based on the aberration theory were proposed for the design of concave gratings. A flat-field concave grating with dimensions of 1 mm × 4 mm× 3.7 mm was designed, achieving spectral resolution of 2nm at 900nm wavelength. A concave grating was fabricated on a plano-concave lens substrate using a custom low-cost holographic technology. Characterizations of the concave grating proved its light wavelength separation and focusing properties.
2. A time-gated CMOS TG-SPAD front-end was designed, with a fixed gate window of 3.5ns. The chip was characterized at gating frequencies up to 100MHz, achieving reduced dark count probability and negligible afterpulsing probability (<1%) for hold-off times longer than 16ns. Temperature dependence of the dark count probability of the TG-SPAD was measured. Activation energy of generation was extracted. The low value of activation energy implies an increased trap assisted tunnelling for SPADs fabricated by deep submicron CMOS technology.
3. A prototype of the time-gated spectrometer was built with the custom concave grating and the time-gated CMOS SPAD. Measured with a 7ps pulsed laser, the performance characterization of the prototype demonstrated that better than 60ps temporal resolution is achieved at 532nm wavelength.
4. Raman peaks of Rhodamine B were resolved by the time-gated spectrometer, measured with different detection windows from 250ps to 3ns to suppress the strong background fluorescence. Lifetimes of Rhodamine B and Rhodamine 6G were extracted through the time-gated fluorescence lifetime measurement. The measured values of 1.52ns and 3.94ns were close to the reference values of 1.68ns and 4.08ns.

Publications:

1. Z. Li, M. J. Deen, S. Kumar, and P. R. Selvaganapathy, "Raman Spectroscopy for In-Line Water Quality Monitoring—Instrumentation and Potential," *Sensors***,** vol. **14**(9), pp. 17275-17303 (2014).
2. Z. Li and M. J. Deen, "Towards a portable Raman spectrometer using a concave grating and a time-gated CMOS SPAD," *Optics Express*, vol. **22**(15), pp. 18736-18747 (2014).
3. Z. Li, M. J. Deen, Q. Fang, and P. Selvaganapathy, "Design of a flat field concave-grating-based micro-Raman spectrometer for environmental applications," *Appl Opt***, vol. 51**(28), pp. 6855-6863 (2012).
4. Z. Li, M. J. Deen, R. Selvaganapathy and Q. Fang. "Single Photon Avalanche Diode for a Time-Gated Raman Spectrometer," *The Electrochemical Society* 225th *Meeting*, Orlando, USA, p. 1493 (11-15May 2014).
5. H. Alhemsi, Z. Li and M. J. Deen. "Time-resolved near-infrared spectroscopic imaging systems," *Second Saudi International Electronics, Communications and Photonics Conference* (SIECPC 2013), Riyadh, Saudi Arabia, 6 pages (27-30, April 2013).

Chapter 2

Raman Spectroscopy and System Design

Raman spectroscopy is a powerful technique for characterization of the chemical composition. To design a portable Raman system for field applications, the basic principle of Raman spectroscopy and its instrumentation are studied. In this chapter, reviews of the theory of Raman spectroscopy, instrumentation development, and advanced Raman techniques are presented in sections 2.1, 2.2, and 2.3, respectively. The system design of the portable Raman system is then given in section 2.4, followed by a summary in section 2.5.

2.1 Raman spectroscopy

2.1.1 Scattering Theory

Raman scattering originates from the interaction between incident electromagnetic radiation and molecular vibrations. In Figure 2.1, the energy level diagram of IR absorption, scattering and fluorescence emission are shown. Bold horizontal lines represent the limits of the bands (E0, E1…) of electronic energy states. Within each electronic energy state band, there are multiple vibrational energy states (gray straight lines). In a normal Raman scattering process, a molecule is excited from its electronic ground state (E0) to a virtual state. This virtual state can be considered as a very short-lived distortion of the electron cloud caused by the incident photon [26]. For Stokes Raman scattering, the initial state of a molecule is the lowest ground state (E0, $v = 0$). In contrast, the anti-Stokes Raman scattering starts from the higher vibrational energy level (E0, $v = 1$) of the ground state [33]. According to the Maxwell-Boltzmann distribution law, the population of molecules at $v = 0$ is higher than that at $v = 1$, so the Stokes Raman scattering is stronger than the anti-Stokes Raman scattering under normal conditions [34]. Hence, the Stokes spectrum is mainly measured by the majority of commercial Raman spectrometers. In addition to the Raman scattering, Rayleigh

scattering also occurs, which has the same frequency as the excitation photon, and its intensity is several orders higher than that of the Raman signal.

Although the Raman scattering efficiency is very low, the scattering can be enhanced by 10^6 [16] if the excitation energy is close to the energy of an electronic transition. This is Resonance Raman scattering and is also shown in Figure 2.1. Since the electronic transition frequencies vary among different chemicals, then ideally, a tuneable laser is preferred for Resonance Raman spectroscopy.

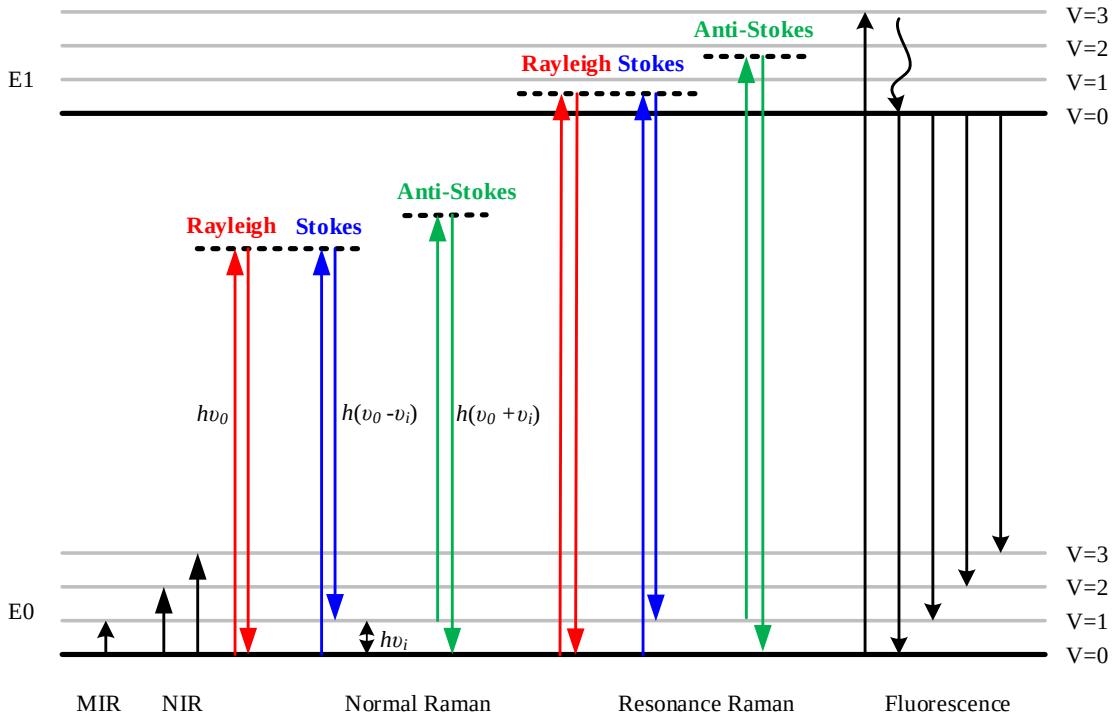


Figure 2.1: Energy level diagram related to IR absorption, Raman scattering and fluorescence emission

The fluorescence emission process is also shown in Figure 2.1. In this process, an electron is excited from the ground electronic state (E0) to the first excited electronic sate (E1) by absorbing a photon. In the excited electronic state, energy can be dissipated through the non-radiative process, depicted as the curved arrow in Figure 2.1. The vibrational relaxation is a non-radiative process which the energy of the electron is given away to other vibrational modes, in a time between 10^{-14} and 10^{-11} seconds [35]. After that, the electron transitions from the excited electronic state to the ground electronic state with the emission of a photon, in a time scale of 10^{-9} to 10^{-7} seconds [35]. The energy of the emitted photon is lower than that of the incident photon because of the energy loss during vibrational relaxation, and this is also known as fluorescence red shift.

The fluorescence emission spectrum is broad band, overlapping the wavelength band of the Stokes Raman signal with selected excitation wavelength. In addition, the intensity of the fluorescence signal is several orders of magnitude higher than that of the Raman signal, so the detection of Raman scattering is very difficult when strong fluorescence emission is present. To suppress or remove the fluorescence background, several techniques have been used. For instance, it is possible to obtain a fluorescence-free Raman spectrum by using a NIR source. Alternatively, since the fluorescence emission is red shifted, the anti-Stokes Raman signal does not overlap with the fluorescence emission spectrum, and the anti-Stokes Raman can be measured in the presence of fluorescence.

2.1.2 Raman Scattering Intensity

A Raman spectrometer measures the intensity of a Raman signal (Stokes or anti-Stokes) and plots the Raman signal intensity versus the frequency shift of the Raman signal relative to the excitation source, known as the Raman shift. As shown in Figure 2.1, both Raman shift and IR absorption are related to the fundamental vibrational modes. Since most of the fundamental transitions occur at MIR, MIR sources are mainly used for absorption spectroscopy. Unlike absorption spectroscopy, sources varying from the ultraviolet (UV) to visible or even NIR regions are capable of Raman scattering excitation. The intensity of Raman signal I_R is wavelength dependent and can be expressed as [26], [36]

$$I_R \propto I_0 \left(\nu_0 \pm \nu_i\right)^4, \tag{2.1}$$

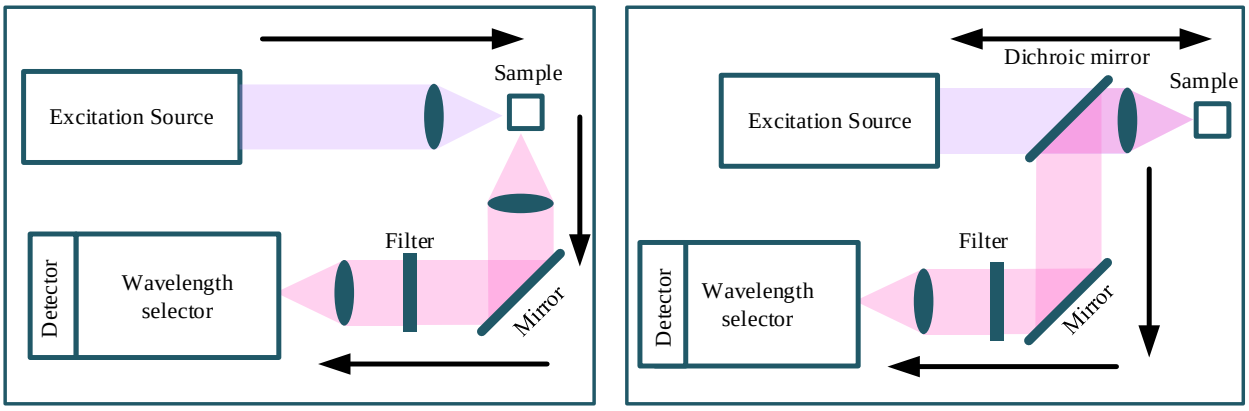
where I_0 is the intensity of the excitation source. The Raman signal intensity I_R in Eq. (2.1) is dependent on the 4^{th} power of the excitation frequency. To achieve a high Raman scattering efficiency, high frequency or short wavelength excitations, such as with a UV source, are usually preferred. However, most modern Raman instruments are equipped with a NIR source or a visible source. UV sources are rarely used due to the unavailability of low-cost UV lasers. In addition, the use of UV or even visible sources excitation could induce a fluorescence signal.

2.2 Instrumentation for Raman Spectroscopy

Generally, a Raman spectrometer consists of four components: an excitation source; illumination and light collection optics; a wavelength selector unit and a detector, as shown in Figure 2.2. There are two types of light collection systems — 90° and 180° configurations [34]. In the 90°

configuration, the scattering light is collected from a direction perpendicular to the excitation direction.

Alternatively, in the 180° configuration, the scattered Raman signal is collected in the direction opposite to the direction propagation of the excitation, which is also termed as back scattering. Additional optics is present in the 180° configuration, such as a dichroic mirror, which transmits the excitation signal and reflects the longer-wavelength scattered signal. As can be seen in Figure 2.2, besides the light collection system, the other three elements (excitation source, wavelength selector and detector) in a Raman spectrometer are the same, and are discussed in the following subsections.



(a): 90 deg configuration (b): 180 deg configuration

Figure 2.2: System with (a) 90°; (b) 180° configurations for collection of the Raman scattering

2.2.1 Excitation source

The bandwidth and power of the excitation source play important roles in the performance of a Raman spectrometer, affecting the resolution for Raman spectra. As mentioned above, the frequency shift between the Raman signal and the excitation signal is related to the fundamental vibrational frequency of a molecule, and the shifts of some vibrations are relatively small. Therefore, a highly-monochromatic and stable beam with high power is necessary for reliable acquisition of the Raman spectrum.

Raman spectroscopy was a niche technique in the 1940s and 1950s primarily due to the lack of powerful excitation sources. Initially, mercury lamps with filters that transmitted a narrow wavelength band were used as the excitation sources, but the beam intensities were low. Raman spectroscopy gained mainstream significance with the advent of lasers in the 1960s. A continuous HeNe red-light laser was first used in a Raman spectrometer in 1963 [37]. The development of

other gas lasers such as Argon and Krypton lasers in the 1970s enabled the application of Raman spectroscopy in the visible and UV regions. In the 1980s, the advent of solid-state lasers (YAG laser, 1064nm) boosted the development of Raman spectroscopy in NIR. The technique of Fourier Transform (FT) Raman spectroscopy [38], [39] became feasible in 1986. FT-Raman spectroscopy using a NIR source could provide fluorescence free spectra, owing to the low energy of a NIR photon, which has marginal probability to bring the molecule from the ground electronic state into the excited electronic state. Many recent continuous wave and time-resolved Raman spectrometers use diode lasers as excitation sources with wavelengths ranging from blue light to the NIR. Table 2.1 provides a list of commonly used lasers in Raman spectrometers for different applications.

Table 2.1: Commercial lasers used in Raman spectrometers and applications

Excitation source	Laser types and wavelength	Techniques of Raman spectroscopy	Applications
NIR source	• Diode laser: 785, 830nm • Solid state laser: Nd-YAG (1064nm), Ti-Sapphire	• FT-Raman spectroscopy • Normal Raman spectroscopy • Surface enhanced-RS	Biological samples Polymers General purpose
Visible source	• Ion laser: He-Ne (633nm), He-Cd (442nm), Ar+ (488nm, 514nm) • Solid state laser: Nd-YAG (532nm), Ti-Sapphire	• Normal Raman spectroscopy • Surface enhanced -RS • Time Resolved Raman spectroscopy • Resonance Raman spectroscopy	Organic components Art, archeology and forensics Semiconductor, minerals General purpose
UV source	• Ion laser: He-Cd (325nm), Ar+ (244nm, 257nm) • Solid state laser pumped dye laser: Ti-Sapphire	• UV Raman spectroscopy • Resonance Raman spectroscopy • Time Resolved Raman spectroscopy	Protein, DNA Natural chromophores Wide bandgap semiconductors

Nd-YAG: neodymium-doped yttrium aluminum garnet; Ti: Titanium; He-Ne: Helium-neon; He-Cd: Helium cadmium; Ar: Argon; FT: Fourier transform; UV: Ultraviolet; DNA: Deoxyribonucleic acid.

The excitation source in a Raman spectrometer is selected to match the requirements of the specific application. The excitation bands of most fluorophores are in UV or visible wavelength regions. Therefore, the excitation with a longer wavelength, such as in the NIR region, greatly weakens the fluorescence signals, owing to the lower photon energy, as explained above. Since natural fluorophores exist in biological samples, including in natural water [40], then NIR lasers are preferred in both dispersive and FT-Raman spectrometers for characterization of biological samples. Furthermore, using NIR source, the photons have lower frequency and energy, and the excitation power can be reduced, which also helps in avoiding damages to the samples.

However, the low photon intensity of a Raman signal in NIR requires expensive optics and detectors [41]-[43]. Therefore, visible-light lasers are usually used for non-biological applications, such as in nanotechnology and solid-state physics. Due to their high stability and low cost, 532nm solid-state lasers are widely used for measurements of inorganic samples. Advantages of visible excitation are the higher signal-to-noise ratio and better sensitivity. In case when both fluorescence

rejection and high sensitivity are required, a compromise is made between these two factors. A common source for biological samples is the 785nm laser, which not only provides some fluorescence rejection, but also works well with silicon-based detectors [44], [45].

2.2.2 Wavelength Selector

The wavelength selector is the most critical component in a Raman spectrometer, through which the information at individual frequencies in the spectrum is resolved. There are basically two types of wavelength selection mechanisms, dispersive and non-dispersive. A dispersive spectrometer relies on spatial separation of wavelengths, using diffraction gratings or prisms. For the non-dispersive spectrometer, light can be selected either by an optical filter or by an interferometer, such as the FT-Raman spectrometer.

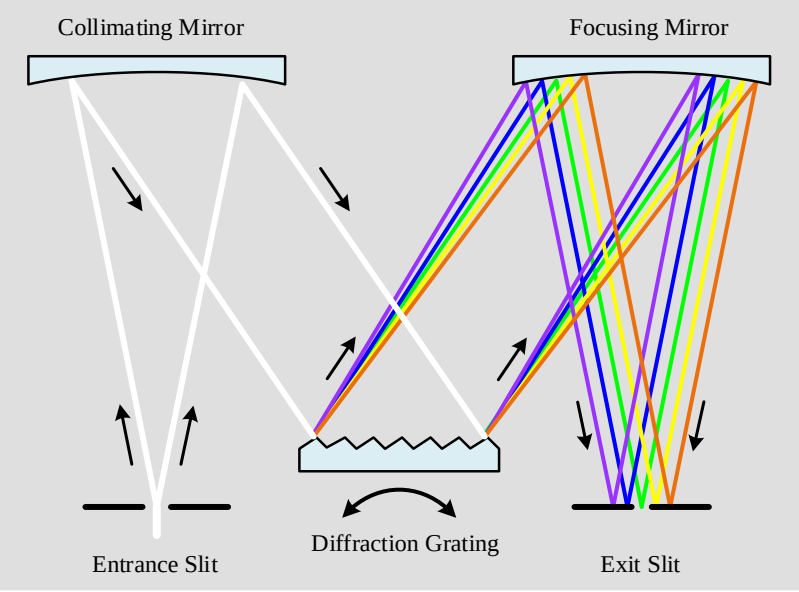


Figure 2.3: Schematic of a monochromator

2.2.2.1 Dispersive Raman Spectrometer

A monochromator is usually used in a dispersive spectrometer, in which there are entrance and exit slits, collimating mirror, diffraction grating, and focusing mirror. Schematic of a monochromator is shown in Figure 2.3. The incident light enters the monochromator through the entrance slit, passing by the collimating mirror. Then, the collimated light is directed to the diffraction grating, which separates the light spatially according to the wavelength. After separation, the dispersed beam from the grating is re-focused on the exit slit by the focusing mirror, so that only the light with the desired wavelength exits from the monochromator.

Two parameters of main importance to a wavelength selector are the wavelength range and spectral resolution. Both parameters depend on the dispersion property of the diffraction grating. For Raman spectroscopy applications, the wavelength range is determined by the excitation wavelength and the Raman shift of the sample. High spectral resolution is desired to resolve the weak Raman peaks.

In addition to the wavelength range and spectral resolution, another issue of importance is the quality of the measured Raman spectrum. As discussed above, the efficiency of normal Raman scattering is very low, while the intensity of Rayleigh scattering is 10^4-10^6 times higher than the intensity of normal Raman scattering [36]. In cases when the Raman spectrum line of the target is very close to the excitation wavelength of the laser, the intensity of stray light from Rayleigh scattering can easily exceed the intensity of Raman signal. To efficiently detect the Raman signal, Rayleigh signals must be rejected or attenuated, and this is known as stray light rejection in a spectrometer. Most commercial spectrometers are equipped with single stage grating and filters for Rayleigh light attenuation and rejection. The development of holographic notch filters and other edge filters has significantly simplified the optical assemblies in Raman spectrometers [46].

Commercial Raman spectrometers are usually equipped with sophisticated but large size components. These spectrometers provide high spectral resolution and throughput, but are expensive. A number of efforts have been made to miniaturize the spectrometers to enable their field usage. For instance, various configurations such as single or double planar gratings [47], single mirror [48] or double mirrors have been proposed for system miniaturization. In addition, many grating designs have been investigated to optimize the diffraction efficiency and spectral resolution [49]. In the past decades, different types of gratings (planar, concave, constant or varied line space gratings) were designed in miniaturized spectrometers.

In addition to the conventional mechanical ruling method, other advanced grating fabrication technologies have been used for fabrication of gratings for different wavelength ranges. As examples, photolithography is usually used to fabricate gratings with a pitch larger than 1μm, the holographic method is used for fine-pitch and aberration corrected gratings [50], and UV nano-imprint lithography [51] and deep X-ray lithography have been used to fabricate concave gratings [52]. In Table 2.2, some of the grating designs for miniaturized spectrometers proposed in the past decade are listed. From Table 2.2, sub-nanometer spectral resolution is achievable for the millimeter-sized spectrometers, which is important for resolving narrow Raman peaks.

Table 2.2: Miniaturized spectrometers with nanometer spectrum resolution

Wavelength range (nm)	Grating Type and Pitch	Diffraction order	Grating Size	Spectral Resolution	Throughput/ Numerical Aperture	Ref
450-750	Double Planar, 1μm	-1	3x3x11mm^3	3nm	9%	[47]
420-770	Planar, 1.6μm	1	0.5cm^3	0.7nm	0.22	[48]
600-700	Planar, 2μm	1	11x1.5x3mm^3	6nm	0.05rad	[53]
400-1030	Concave, 3.2-4μm	Multi-order	11x6x5mm^3	2.5nm	0.2	[50]
580-730	Concave, 4μm		R=25.8mm	0.9nm	0.11	[54]
1475-1625	Concave, 3μm	3	R=44.4mm	1.1nm	0.21	[52]
512-768	Concave, 6μm	2	30x30x2mm^3	2.8nm		[55]

2.2.2.2 FT-Raman Spectrometer

Although the feasibility of FT-Raman was demonstrated as early as 1964 [56], due to technology limitations, the first realization of FT-Raman spectrometer came much later, in 1986 [38]. Differing from the dispersive Raman spectrometer, a FT-Raman spectrometer uses an interferometer for wavelength separation. The measured signal is an interferogram in time domain, from which the Raman spectrum is obtained after a Fourier transformation.

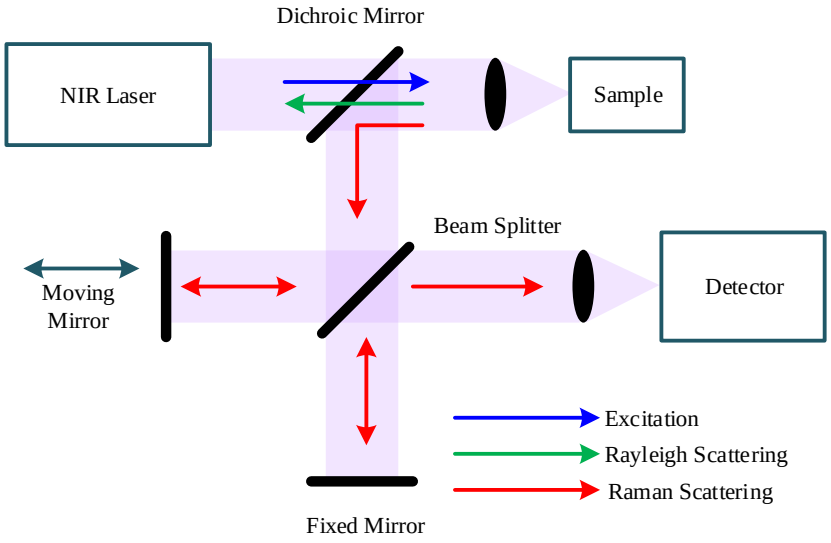


Figure 2.4: Simplified block diagram of the FT-Raman spectrometer

Figure 2.4 illustrates a FT-Raman spectrometer based on a Michelson interferometer. The excitation source (e.g., from a NIR laser) is directed to the sample through a dichroic mirror and lens, the latter also collecting the scattered signals (Rayleigh and Raman) from the sample. When reaching the dichroic mirror, the short-wavelength Rayleigh scattering signal is transmitted and the

longer wavelength Raman scattering signal is reflected to a beam splitter. Through the beam splitter, half of the Raman signal is transmitted to a fixed mirror and the other half is reflected to a moving mirror. Because of the optical path difference caused by the moving mirror, the two beams reflected from the two mirrors undergo constructive or destructive interference. Finally, the signal is registered on the detector and the interferogram can be obtained from the detector.

Similar to the dispersive Raman spectrometer, spectral resolution is important for a FT-Raman spectrometer. Spectral resolution (R) of a FT-Raman spectrometer is determined by the maximum travel range (Δx_{max}) of the moving mirror [16], which can be written as

$$R = 1/\Delta x_{max}. \tag{2.2}$$

If the maximum travel range of the moving mirror is 1cm, then the spectral resolution of the spectrometer is $1cm^{-1}$. Generally, different spectral resolutions can be selected in a commercial FT-Raman spectrometer. The second parameter of great importance is the wavelength range, and it is related to the type of material of the beam splitter. In addition, the cut-off wavelength of the NIR detector also affects the wavelength range, and this will be discussed later in section 2.2.3.

Compared with the dispersive Raman spectrometer, the FT-Raman spectrometer has a higher throughput, excellent frequency accuracy and precision, and higher resolution. Moreover, owing to the use of a NIR excitation source (1064nm), fluorescence emission can be suppressed. However, the use of a longer wavelength excitation has several limitations. First, NIR absorption spectroscopy occurs in this region, which attenuates the incident light. Second, the intensity of the Raman signal is proportional to the 4^{th} power of the incident frequency. As a consequence, Raman scattering efficiency for the longer wavelengths of NIR is significantly lower than that for the shorter wavelengths of the visible light. The lower Raman scattering efficiency limits the sensitivity of the FT-Raman spectrometer, which is important in applications such as those to detect water contaminants. Therefore, FT-Raman spectrometers are mainly used only when the samples' fluorescence is high, such as in forensic analysis [57] and pharmaceutical applications [58].

2.2.3 Detector

Because of the low Raman scattering efficiency, detection of Raman signals is very challenging, and the detector must be very sensitive. A detector exploits the photoelectric effect which uses the energy of the incoming photons to generate charge carriers that are separated and subsequently measured as a current at the terminals [59]. Key parameters associated with a detector are the

quantum efficiency (QE) and the dark current. QE defines the efficiency of a detector to convert optical photons to photon current and dark current refers to the current caused by the non-photon generated charge carriers. Accordingly, to observe the weak Raman signal, the detector should have high QE in the related wavelength band, low dark current and wide dynamic range. To date, several types of detectors have been successfully used in Raman spectrometers, and many of them are discussed in the following subsections.

2.2.3.1 Photomultiplier Tubes (PMT)

A PMT consists of a photocathode, series of dynodes and an anode. Photons incident on the cathode generate electrons due to the photoelectric effect. These electrons are accelerated by the high electric field between the cathode and an adjacent dynode. The accelerated electrons impinge on the dynode, generating additional electrons due to secondary emission. A cascade structure of such dynodes quickly multiplies the number of electrons, which generate a large current pulse when reaching the anode. Because of the high gain of charge multiplication, a PMT can detect even a single photon and thus normally works in the photon counting mode.

In comparison with other types of detectors, advantages of PMTs are their high gain and short transit time. Modern PMTs have gain of above 10^5, dark current in the range of nA, and transit time is in the range of nanoseconds [60]. Their main disadvantages are that high operating voltages are required for the high gain. The operation voltage in commercial PMTs [61] is typically above 1000V. The QE of PMTs in the visible to NIR region is below 40% [61] and is lower than the QE of commercial charge-coupled devices (CCDs). Because of their large active region (~10mm), PMTs are mainly used in monochromators, and continuous wavelength measurement is realized by scanning the gratings to adjust the output wavelengths. PMTs were widely used in dispersive Raman spectrometers before the 1980s because of their high QE and low dark current. However, with the advent of high QE multichannel detectors in the 1980s (CCDs), PMT is less used in modern Raman spectrometers.

2.2.3.2 Charge coupled devices (CCDs)

CCDs consist of a large matrix of pixel elements, the fundamental structure of which is a metal-oxide-semiconductor diode on a thin silicon substrate. A polysilicon gate is deposited on top of each pixel, and an external bias is applied to the gate to control the potential of the silicon region beneath the gate. By applying different biases (reverse bias and zero), an individual pixel is isolated from the neighbouring pixels because of the insulating barriers around the potential well shown in

Figure 2.5. For pixels that are reverse-biased, depletion regions are formed, and charges are held and stored within the potential well up to the full well when illumination is applied. In contrast, those zero-biased pixels are transparent to the incoming photons. The number of charges generated is proportional to the intensity of the incident light flux, and the full well capacity determines the maximum light intensity that can be detected. By adjusting the bias, carriers stored in the potential well can be transferred to the output of the detector.

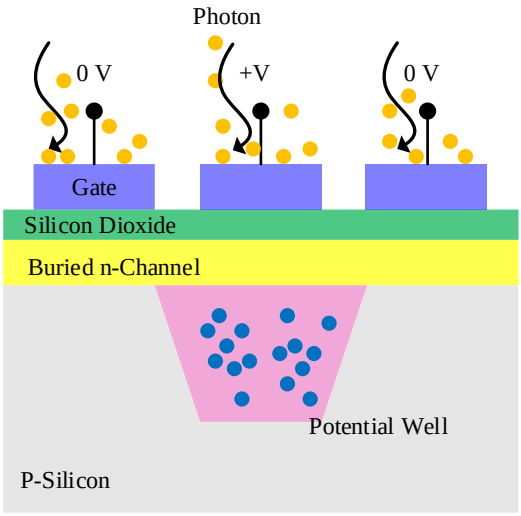


Figure 2.5: Basic structure of a CCD

The active region of a pixel in a CCD is typically 10 µm, a small size that allows for on-chip integration of pixel arrays for multichannel detection systems. Multichannel detection allows for the detection of multiple wavelengths simultaneously. It is beneficial to reduce the integration time and the risk to damage samples when using long exposure times [16].

QE is defined as the probability that a single photon incident on the device will generate an electron-hole pair that contributes to the photon current. Since not all photons are absorbed and produce electron-hole pairs, QE can therefore be written as [62]:

$$\mathrm{QE} = (1-r)\cdot\zeta\cdot\mathrm{PAE} = (1-r)\cdot\zeta\cdot\left[1-\exp(-\alpha W)\right], \tag{2.3}$$

where r is the surface reflectance of the optical power, ζ refers to the fraction of electron-hole pairs that contribute to the photon current, α is the absorption coefficient of the material, and W is the width of depletion region.

1) The first term (1-*r*) in Eq. (2.3) is associated with the reflectance occurring at the surface of the detector. Not all photons can pass the surface layers of the detector. To reduce surface reflectance and improve QE, antireflection coatings can be used.
2) The second term ζ in Eq. (2.3) considers the recombination effect of photon generated electron-hole pairs.
3) The third term in Eq. (2.3) refers to the photon absorption efficiency (PAE) in the detector. Determined by the different absorption coefficients of photons with different wavelengths in material, fraction of the incident photons in certain wavelength range can be absorbed within the depletion region of the detector. The rest photons might be absorbed out of the depletion region or pass the detector without absorption. To improve QE, it is important for the detector to have sufficiently large depletion region width *W*.

Mainstream CCDs are silicon based and have a peak QE of above 90% and the maximum detectable wavelength is ~1100nm (due to the band gap of silicon – 1.12 eV). Hence, CCDs have become the mainstream detectors for commercial multichannel spectrometers (Horiba Jobin Yvon, Kyoto, Japan) in the visible region. However, the QE of most standard silicon based CCDs is limited by the available width of the depletion region, which decreases rapidly when the wavelength is beyond 900nm, or when the depletion width is less than 5μm, as shown in Fig. 2.6.

Raman shifts of most bacterial substances and chemicals are between 500cm^{-1} and 3000cm^{-1} [63], [64]. The corresponding Stokes Raman signals under 785nm excitation are between 817-1026nm, which covers the lower QE region of standard CCDs. To measure the full Raman spectrum, novel designs have been implemented in modern advanced CCDs, by modifying either the position or width of the depletion region. Deep depletion and back illumination are two typical techniques used in industry to increase the QE at longer wavelengths (Princeton Instruments, Andor Technology) [65]. Moreover, other techniques have also been developed to increase the detection efficiency for ultra-low light level detection, including the intensified CCD (ICCD) and Electron-Multiplying CCD (EMCCD). With respect to dark current, thermal generation is the main source of the dark current. Hence, cooling is a direct and efficient way used in commercial CCDs to reduce the dark current. Methods such as cryogenic cooling with liquid nitrogen and thermoelectric cooling have been used.

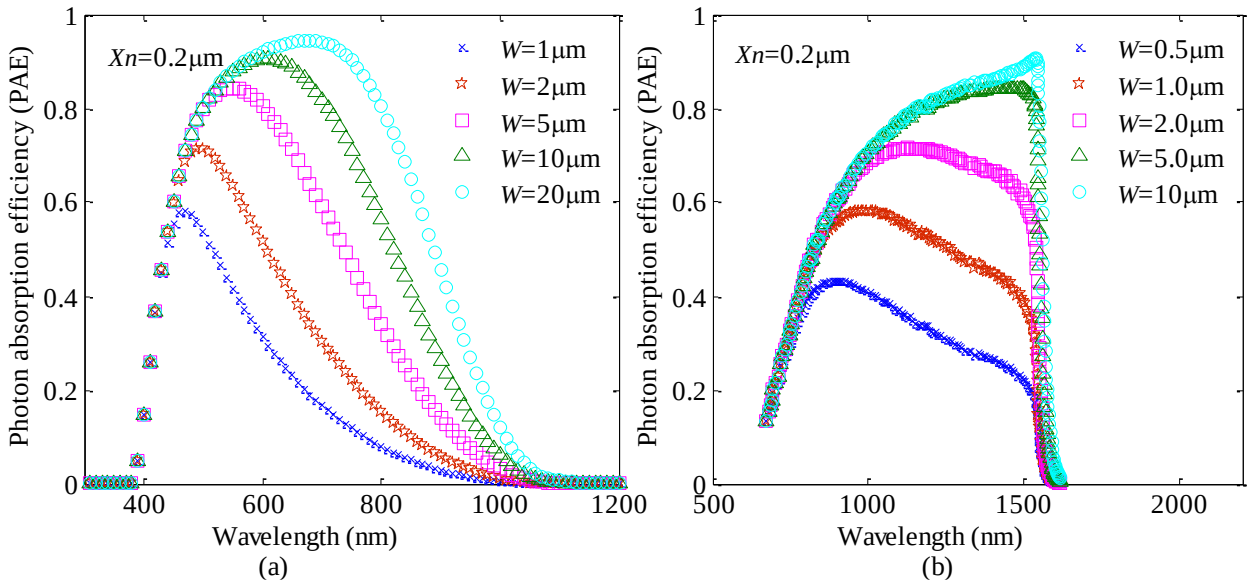


Figure 2.6: Photon absorption efficiency with different wavelength: (a) Si; (b) Ge. (Xn: distance from surface to depletion region; W: depletion region width)

2.2.3.3 Silicon Avalanche Photodiode

The silicon-based avalanche photodiode (APD) is a PN junction working in the reverse-biased mode. When incident photons are absorbed, electron-hole pairs are generated in the depletion region and multiplied through the avalanche multiplication process. APDs are sensitive detectors with internal gain, although their gain is lower than PMTs. Similar to CCDs, the lower QE of the photodiodes at longer wavelength is determined by the energy band gap of silicon. Increasing the width of the depletion region or using a different material with a narrow band gap can improve the detection efficiency at longer wavelengths. This dependence can be observed through the simulation of the photon absorption efficiency (PAE), as given in Eq. (2.3). Since the depletion region is not at the surface of the material, PAE is modified as:

$$\mathrm{PAE}=\frac{\int_{X_n}^{X_n+W}\alpha(\lambda)P_0\exp\left[-\alpha(\lambda)x\right]dx}{\int_0^{\infty}\alpha(\lambda)P_0\exp\left[-\alpha(\lambda)x\right]dx}=\left(1-\exp\left[-\alpha(\lambda)W\right]\right)\exp\left[-\alpha(\lambda)X_n\right], \tag{2.4}$$

assuming the distance from the surface of the detector to the depletion region is X_n. α(λ) in Eq. (2.3) is the absorption coefficient of light in the material [66], which depends on the wavelength of the incident light. Based on Eq. (2.4), Figure 2.6 shows the simulation results of the dependence of PAE on the width of the depletion region and its distance from the surface of the detector for both silicon and germanium (Ge). From the simulation results, we can see that the silicon-based detector has high PAE in visible and short NIR regions, while the Ge-based detector has high PAE at longer wavelengths up to 1500nm.

The silicon-based APD is robust, inexpensive, and easy to miniaturize and fabricate in standard semiconductor technologies. It can be operated at a lower voltage supply compared to PMTs, and is compatible with CMOS control circuitry. Although CCD is the common detector in commercial Raman instruments, APD is also a promising technology for portable and high-speed detection systems. In past decades, significant progress has been achieved for APDs through improvements in gain, size of active region, and response time. Commercial APDs are available from a variety of companies, for example, Hamamatsu Corporation, Boston Electronics and OSI Optoelectronics. In academia, an APD with 7.6nA/mm^2 dark current density, fabricated by CMOS 0.35μm technology, was reported in 2008 [67]. Another APD with a gain of 569 and 3.2GHz 3dB bandwidth under 10.6V reverse bias was described in [68].

The single photon avalanche diode (SPAD) is essentially a PN junction biased above the avalanche breakdown voltage. The avalanche charge multiplication of photon-generated electron-hole pair by single photon produces pulses as in PMT. Counting these pulses is equivalent of counting photons, and this method of use of APD is also known as the Geiger mode. In a SPAD, single photon generated free carriers are multiplied by impact ionization in the very high-electric field of the depletion region, which then triggers the self-sustaining avalanche process. The voltage pulse is registered by a readout circuit, with the leading edge marking the arrival time of a photon. Owing to its single photon sensitivity, the SPAD always works in the photon counting mode. Similar to other photodetectors, the thermally generated dark count is an important issue for the application of SPADs. The past decade has witnessed the development of SPADs with higher detection efficiency, lower dark count level, higher detection rate, and higher fill factor. Low-cost SPADs have been realized by using inexpensive mainstream CMOS technology as shown in Figure 2.7. The photon detection is fulfilled by the N+/P-well diode, and detailed information on this structure will be given in chapter 4.

2.2.3.4 NIR Detectors

To detect the Raman signal in NIR region (>1μm), an indium gallium arsenide (InGaAs) detector is commonly used. Because of its lower energy bandgap, thermal generation of dark current is strong and InGaAs detectors are usually cooled to liquid nitrogen temperature (77K) to control the thermally generated dark current. However, the cut-off wavelength of the QE shifts to shorter wavelength with more cooling. This wavelength shift is caused by the negative dependence of the bandgap energy on temperature. Take the FT-Raman spectrometer for example, when using a

1064nm laser for excitation, the maximum Raman shift of InGaAs detector drops from 3600cm^{-1} at room temperature to 2900cm^{-1} when the detector is cooled to 77K [69]. In addition, the Germanium (Ge) detector is also a mature detector in the NIR region. Today, both InGaAs and Ge detectors have been used in commercial FT-Raman spectrometers [36].

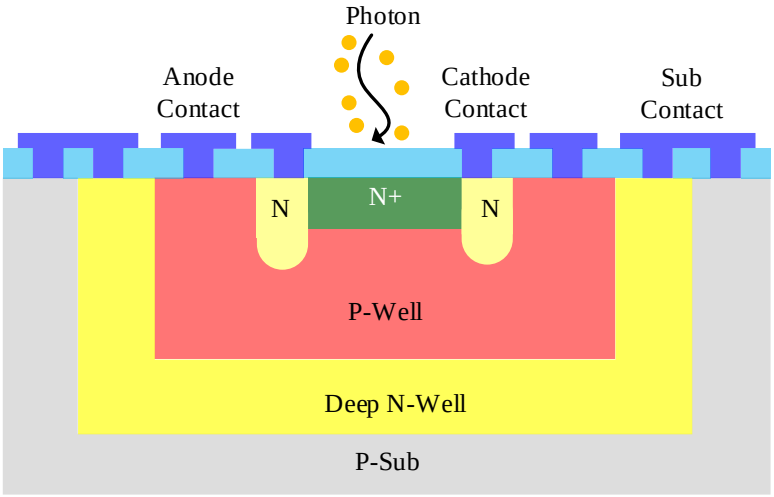


Figure 2.7: Cross section of a CMOS SPAD

2.3 Advanced Raman Techniques

Despite the advantages of Raman spectroscopy, using it for environmental detection of chemical contaminants, such as in water quality monitoring, is difficult due to the low LOD required for this application. The weak Raman spectrum compounds this challenge. To improve LOD and extend its application in the detection of low-concentration samples, various techniques have been developed for Raman spectroscopy. The improvement of LOD can be realized in two ways,

1) enhancing the scattering intensity, and
2) reducing background signal.

The Raman signal can be enhanced by pre-concentrating the contaminants in the sample. On the other hand, the background signal has been reduced by rejection of fluorescence and reduction of the detector dark counts using special design or optimized techniques.

Sample pre-concentration is the most direct strategy to enhance Raman scattering. To date, many pre-concentration techniques have been proposed for both Raman and IR absorption spectroscopy, for example, solid phase micro-extraction (SPME) [70]. Microfluidic devices have been used as miniaturized pre-concentrators of chemicals and contaminants [71], [72]. In [73], PDMS was used with SPME to pre-concentrate organic compounds, and the Raman signal intensity

was reported to be increased by more than two orders of magnitude. However, most of the pre-concentration strategies are time consuming and the experimental setups are complex. Alternatively, advanced techniques have been developed and applied in modern Raman instruments to either increase Raman scattering or further suppress the fluorescence background. Some of these advanced techniques will be discussed in the following subsections.

2.3.1 Surface Enhanced Raman Spectroscopy (SERS)

Currently, surface enhanced Raman spectroscopy (SERS) is the most efficient Raman technique for detection of very low concentration of targets. Since its first observation in 1974, SERS has been widely researched in academia and the number of papers published annually on this topic is growing rapidly. Detailed reviews of SERS including the fundamentals, active substrates and its application can be found in [74]-[76]. Here, a brief discussion of the development of SERS, its instrumentation and application in the detection of low concentration contaminants is given.

2.3.1.1 Theory

SERS was first observed on a roughened silver electrode in 1974, and this phenomenon was explained as a consequence of increased surface area [77]. However, later researches attributed this enhancement to the combination of two mechanisms—Electromagnetic (EM) and Charge Transfer (CT) enhancements. When the incident electromagnetic wave interacts with a roughened metal substrate, the localized surface plasmons are excited, oscillating perpendicular to the metal surface, and amplifying the electromagnetic field near the surface. The field enhancement also amplifies the incident light, enhances the Raman scattering intensity. In addition, in situations when the analyte is chemically bonded on to the surface, electrons can transfer between metal and analyte, which can give rise to additional amplification by chemical enhancement. Schematic of SERS is shown in Figure 2.8.

In comparison with normal Raman spectroscopy, the intensity enhancement in SERS was found to be ~ 10^6 [16]. If combined with a Resonance Raman scattering, even higher enhancement can be achieved (10^8-10^9), and an enhancement of ~10^{14} has been obtained [16]. The signal enhancement significantly improves the LOD, and the advent of single molecule SERS in the 1990s has made SERS a promising technique for pharmaceutical and environmental applications [78], [79].

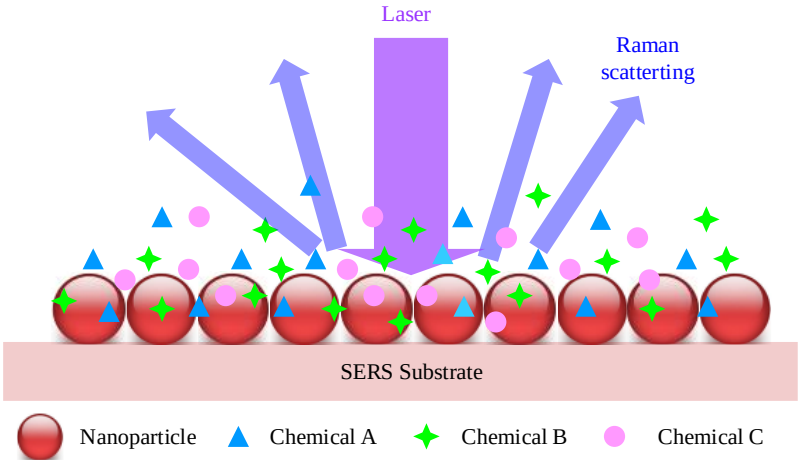


Figure 2.8: Schematic illustration of SERS

2.3.1.2 SERS substrate and fabrication techniques

Although a high enhancement factor can be obtained in SERS, this enhancement is dependent on a variety of factors such as wavelength, substrate profile, and distance [74]. The greatest enhancement is usually observed when using a specialized substrate and when the molecules are absorbed on the surface of the substrate. As a consequence, available fabrication techniques of SERS substrates, play a dominant role in the performance of SERS. Early SERS substrates were electrochemically roughened electrodes. Today, substrates with metallic nanoparticle have been widely used in SERS instruments, building on advances in the development of nanofabrication technologies such as electron beam lithography [80] and nanosphere lithography [74].

Advantages of present nanofabrication are in features to control size, shape and orientation of the nanoparticles, which are of great importance to the intensity enhancement of Raman signal. As mentioned in [75], the wavelength of peak surface plasmon resonance varies with different nanoparticles sizes. Moreover, the type of substrate metals would also affect the enhancement. Silver has been the most commonly used metal for its ability to excite intensive plasmon resonance at visible wavelength, followed by gold [81] and copper. Accordingly, to achieve maximum enhancement for a specific wavelength band, the size of the nanoparticles should be optimized to match the excitation wavelength and also consider the type of metal substrate.

2.3.1.3 SERS in environmental application

The advantage of SERS is the achievable LOD, even in water quality applications. To be applied for water monitoring, LOD of SERS should be lower than the maximum acceptable contaminants

level, such as the maximum contaminants level (MCL) set by the Environmental Protection Agency (EPA). For example, the MCL for cyanide is 200 ppb and MCL for Arsenic is 10 ppb. Table 2.3 lists some of the applications of SERS in water contaminants detection. According to the results in Table 2.3, ultra-low concentration detection is achievable, in particular if combined with microfluidic preconcentration devices [82], [83].

Table 2.3: SERS in water contaminants application

Sample	Laser wavelength, Power, and detection time	SERS Substrate	LOD (M)	Ref
B. subtilis	750,50mW, CCD 1min	AgFON: 600nm diameter	2.1×10^{-14}	[84]
Chromate	785, 80.2mW, CCD TE Cooled 20s	Au/ mercaptoethyl pyridinium	5×10^{-7}	[85]
E. coli	514.5, 100mW, CCD 1-2mins	Ag nanoparticle suspension	~10^3cfu/ml	[86]
Uranium	785, 60mW	Au/aminomethyl phosphonic acid 50-60nm	8×10^{-7}	[87]
RDX in water	785nm, 1mW, CCD 10s	Au nanoparticles 90-100nm	1×10^{-6}	[88]
Mercaptobenzoic Acid	785nm, 5mW, 2s	Ag nanostructure on polyaniline membrane	1×10^{-12}	[89]
Dye molecule	785nm, 2mW, CCD 10s	Fractal-like Au nanostructure 30-50nm	4.3×10^{-9}	[90]
Thrombin	632.8nm, 0.5mW,	Au nanoparticles 56nm	2×10^{-11}	[91]
Arsenite	532nm, 20mW	Ag, Cu nanoparticles coated with poly(vinyl pyrrolidone)	1.3×10^{-8}	[92]
Cyanide	514nm, 20mW, CCD 30s	Ag colloids 35-40nm	$1.5\text{-}2 \times 10^{-8}$	[83]
Malachite green	514nm, 20mW, CCD 30s	Hydroxylamine hydrochloride-reduced Ag colloid 40nm	$2.6\text{-}5.2 \times 10^{-9}$	[82]
Cyanide anions	532nm, 10mW, CCD 100s	Ag nanoparticles immobilized on oxidized silicon substrates	2.7×10^{-7}	[93]
Perchlorate	785nm, 1.5mW, CCD 10s	Ag nanoparticles on functionalized silica sol-gel films	1×10^{-6}	[94]
Polychlorinated biphenyls	532nm, 3.09mW, 30s	AgFON	5×10^{-11}	[95]
Perchlorate	785nm, 1.5mW, CCD 10s	Cystamine-modified Au nanoparticles	5×10^{-16}	[96]
Uranyl Ion	632.8nm, 2mW, CCD 1s	Ag modified polypropylene filter (PPF) substrates	4×10^{-8}	[97]

FON: Film over nanosphere

2.3.2 Time-Gated Raman Spectroscopy

In addition to the enhancement of Raman scattering as in SERS, the signal-to-noise ratio can also be improved by reducing the background signal. Sources of background signal in a Raman measurement can be either from the detector or from the incident optical signal. As mentioned in

section 2.2.3, the thermally generated dark current is an important issue for both CCDs and APDs, and an efficient way to reduce the dark current is to cool the detector during measurements.

With respect to the background signals from the incoming photons, they can be from both Rayleigh scattering and fluorescence emission. Since Raman scattering and Rayleigh scattering differ in their emission wavelengths, so the influence of Rayleigh scattering can be removed using an optical filter. In contrast, the fluorescence emission band overlaps with the Raman peak for certain excitation wavelength, which blurs the Raman peaks. Overall, reducing background fluorescence and detector dark current will lead to a higher signal-to-noise ratio.

Considering temporal distributions, Raman scattering is an instantaneous response to the excitation source, while fluorescence is emitted with a temporal distribution characterized by the so called fluorescence lifetime. The fluorescence lifetime varies from hundreds of picoseconds to tens of nanoseconds depending on the type of samples. Consequently, if a short pulsed excitation source is used, Raman scattering and fluorescence emission can be separated in time domain (Figure 2.9). To realize this separation, the width of the laser pulse should be narrow, usually in the range of hundreds of picoseconds. The repetition rate is selected according to the fluorescence lifetime, which ensures that the fluorescence signal from the previous cycle has fully decayed and does not contribute to the next cycle.

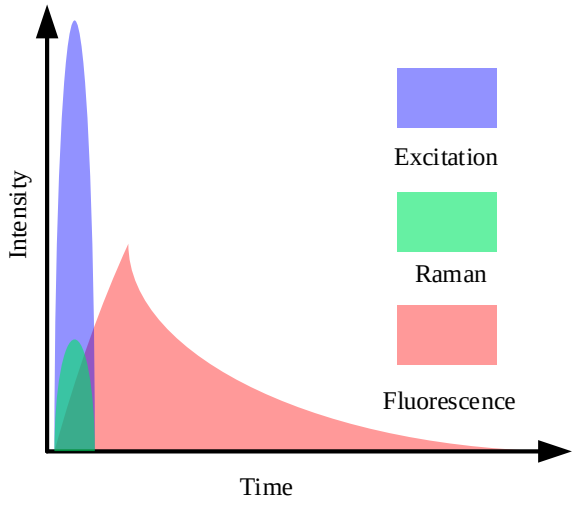


Figure 2.9: Temporal variations of excitation, Raman scattering and fluorescence emission

In time-gated Raman spectroscopy, the detection must be synchronized with the excitation source, so that only the signal overlapping with the detection window in time domain can be detected. This is known as time-gated detection. Two techniques have been used for time-gated

measurements. One technique is a time-gated modulation of the incoming optical signal, by introducing an optical shutter. The other technique is to modulate the detector sensitivity by gating the operation of the detector only in a predetermined short time window, and turning off the detector after this time window.

2.3.2.1 Kerr Gated Raman System

The Kerr gate is the best-known and fastest optical shutter for time-gated Raman spectrometers. A Kerr gate with 25ps response and high repetition rate was proposed as early as the 1970s [98]. Later researches were focused on further reducing the response time. In 1999 [99], the Kerr gate was first introduced to Time Resolved Resonance Raman spectroscopy (TR^3) with a response time of ~3ps. This Kerr gate system consisted of two crossed polarizers, and a Kerr medium was placed between the polarizers. A gating pulse was used to turn on and off the Kerr gate, by varying the polarization orientation of the light passing through the Kerr medium. Otherwise, no light could pass the Kerr gate due to the crossed polarizers. Schematic of the Kerr gate system is shown in Figure 2.10.

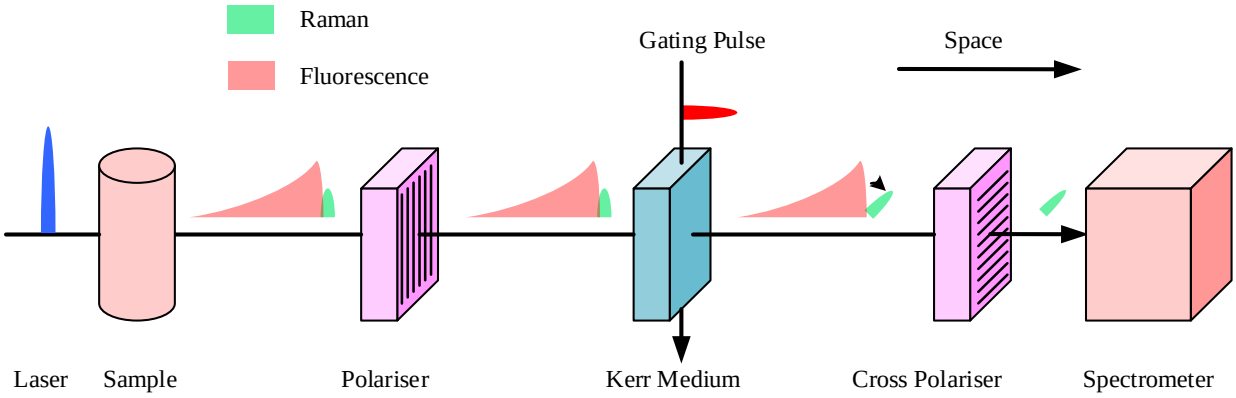


Figure 2.10: Schematic of a Kerr gate system

In the Kerr gate system, if the short gating pulse overlaps in time with the Raman signal, then the Raman signal would be able to pass the gate with the fluorescence signal being blocked. This setup has also been used for a variety of applications including plant auto-fluorescence, depth profiling spectra through the prostate gland and the bladder, and depth profiling of calcifications in breast tissue [100]-[102].

Currently, the Kerr gate has become a very popular optical shutter in time-gated Raman detection. However, owing to its fast response, Kerr gate has been mainly used in time-resolved Raman spectroscopy which aims at analyzing the dynamic response of biology samples. For field

applications, the fast response provided by the Kerr gate is sufficient to perform the function of fluorescence rejection, but the complex setup has limited the Kerr gate to laboratory use.

2.3.2.2 Fast Time-Gated Raman Systems

In addition to the optical shutter, fast gated detectors have also been employed in Raman spectrometers for fluorescence rejection. The most commonly used detector in a time-gated Raman system is the intensified CCD (ICCD). Different from normal CCDs and EMCCDs, an ICCD can be operated in the time-gated mode and perform ultra-sensitive detection down to a single photon. In an ICCD, a gain voltage controlled image intensifier tube is positioned in front of the CCD, and incident photons are multiplied inside the intensifier before being focused onto a CCD. The gain voltage not only determines the multiplication, but also can gate on and off an ICCD. The ICCD has been used as an alternative technique to the Kerr gate system. Although not as fast as the Kerr gate system, most modern ICCDs can achieve hundreds picoseconds gating width, which is adequate for normal Raman spectroscopy [103]-[105]. In Table 2.4, a summary of time-gated Raman system is presented. It is worth noting that a short gate window is achievable for both ICCD and PMT.

Table 2.4 Summary of characteristics of recently developed time-gated Raman systems

Sample	Source	Power	Detector	Detection window	Ref
Rhodamine 6G	532nm, 6.4kHz, 900ps	3μJ/pulse	PMT	700ps	[106]
Explosives	532nm, 50ps	15mJ	ICCD	500ps	[104]
PMMA	398nm, 76MHz, 3ps		ICCD	250ps	[105]
Explosives	532nm, 10kHz, 5ns	140mJ/pulse	ICCD	1us	[107]
Pyrene/toluene Phenylacetone monooxygenase	257nm, 76MHz, 3ps 405nm, 76MHz, 3ps	2mW 10mW	ICCD	300ps	[103]

2.3.3 State-of-the-Art Portable Raman Spectrometers and Design Challenges

For purpose of field applications, many portable Raman spectrometers have been developed in industry. Comparing with bench-top Raman spectrometers, the portable Raman spectrometers are low cost, light weight, and more compact. Table 2.5 lists several portable Raman spectrometers and their specifications. These spectrometers can be battery powered with several hours operation time and fast acquisition times (TruScan RM and NOVA). The 785 nm laser is widely used in these instruments for general purpose applications, and longer wavelength excitation is used when strong

fluorescence is present during measurements (Inspector 500 and CBEx™ 1064). These instruments provide wide spectrum range with ~10 cm^{-1} spectral resolution, and can be used for raw material identification or manufacturing process material validation.

Two constraints restricting the field applications of a Raman spectrometer are size and cost. To design a system targeting field applications, both size and cost must be reduced. As listed in Table 2.5, miniaturized Raman spectrometers are available commercially, from SnRI and ICx. Unfortunately, these portable Raman spectrometers are still very expensive, though they are much cheaper than bench-top systems. In addition, comparing the two models from SnRI, the one with fluorescence suppression (SnRI CBEx^{TM} 1064) is twice the price of the one for general purposes (SnRI CBEx^{TM}). Thus, design challenges of this work arise from building a low-cost portable Raman spectrometer with fluorescence suppression function.

Table 2.5. Commercially developed portable Raman spectrometers

Portable Raman	Excitation Source	Spectral Resolution	Spectrum Range (cm^{-1})	Weight (kg) Size (cm)	Price
Thermal Scientific FirstDefender RM	250-125-75mW	7-10.5cm^{-1}	250-2875	0.816 19.3 x 10.7x 4.4	$50,000
SciAps Inspector-300	300mW @785 nm	6-8cm^{-1}	175-2875	1.70 19.1 x 17.5 x 4.3	
Smiths Detection RespondeR RCI		12cm^{-1}	225-2400	3.1 22 x 9.9 x 19	$30,000
SnRI CBEx^{TM}	100-70mW @Visible		400-2300	0.333 9.1 x 7.1 x 3.8	$15,000
SnRI CBEx^{TM} 1064	300-400mW @1064 nm		400-2300	0.771 11.4 x 7.9 x 5.7	$30,000
ICx Fido Verdict	75mW @785nm	12cm^{-1}	300-2000	0.415 8.4 x 4 x 19	$18,798
Rigaku FirstGuard	30-490mW @1064 nm	15-18cm^{-1}	200-2000	2.30 12.2 x 31.1 x 31.4	$29,360

Among the four components of a Raman spectrometer, the most expensive components are the excitation source and the detector. In contrast, the wavelength selector is not very expensive, but it accounts for the large system size. To overcome the design challenges, the proposed system will be optimized for small size of the wavelength selector and for cost reduction, using a detector fabricated in a low-cost standard CMOS technology.

2.4 System Design

2.4.1 Main System Components

2.4.1.1 Wavelength Selector

The wavelength selector determines the spectral resolution of a Raman spectrometer, and it is also the largest component of the system. To reduce the overall system dimensions, the size of the wavelength selector has to be reduced.

As mentioned in section 2.2.2, monochromators are widely used for wavelength selection in commercial single channel spectrometers. Typically, a monochromator is based on a planar diffraction grating. As the most important parameter of a high spectral resolution spectrometer, the spectral resolution is determined not only by the diffraction limit of the planar grating, but also by other system parameters, such as the numerical aperture of the entrance slit and focusing length of the mirrors. To achieve high spectral resolution, the diffraction grating and mirrors are specially designed and assembled. The complex assembly and the optical paths between lenses, grating, and slits determine the dimension of a monochromator.

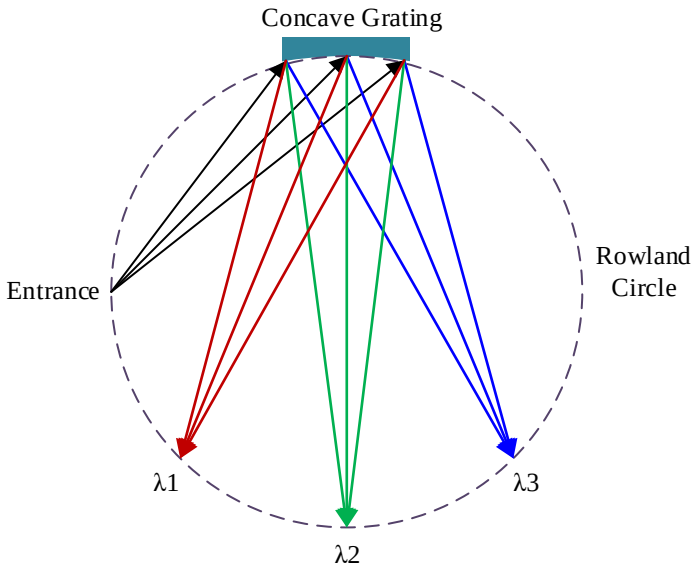


Figure 2.11: Optical diagram of a concave grating based system

In addition, concave gratings are also used in commercial spectrometers. Unlike a planar grating, a concave grating can perform both functions of wavelength separation and light focusing without using extra optical components. Figure 2.11 shows the schematic diagram of a concave grating based Rowland mount system. The Rowland circle in this configuration is a virtual circle,

which is internal tangent to the curvature of the concave grating and its diameter is equal to the radius of the concave grating.

In the Rowland configuration, if the entrance slit is positioned on the Rowland circle, the separated diffraction beams reflected from the grating will also be focused onto the Rowland circle, but at different positions according to the wavelength. The spectrum can be acquired by placing a detector at the focusing point of the Rowland circle. Since only a grating is required, the concave grating based system is more compact and the size of the Raman system can be significantly reduced. Detailed information about the design, fabrication, and characterization of the concave grating will be discussed in chapter 3.

2.4.1.2 Detector

Detection of the Raman signal is the most challenging part of this work, not only due to the low Raman scattering efficiency, but also due to the strong fluorescence signal potentially overlapping the wavelength range of the Raman signal.

As demonstrated in section 2.2.1, a common way to obtain a fluorescence suppressed Raman signal, while ensuring a high enough scattering efficiency, is to use a NIR excitation source, such as a 785nm laser. Unfortunately, the wavelength of the Stokes Raman signal under this excitation covers the low QE region of regular Si-based detectors. Special process-treated Si detectors for longer wavelength applications are commercially available, such as the back illumination CCDs, but they are very expensive. SERS is an effective way to improve the Raman scattering efficiency with the NIR excitation, but the use of SERS substrates increases the cost and not all chemicals are SERS active.

To reduce the cost of the system, Si-based CMOS detectors are the preferred choice. Compared with CCDs, CMOS detectors feature low power consumption, low-cost, high speed, high integration capability, and have been widely used in biomedical applications of low light level detections [108]-[111]. Because the optimum wavelength response of CMOS detectors is usually in visible region, then a visible excitation source is preferred. The short wavelength excitation solves the problem of low Raman scattering efficiency, but it introduces background fluorescence for certain samples. Thus, the detector should not only be very sensitive, but also perform the function of fluorescence suppression.

Considering the different temporal distributions of Raman scattering and fluorescence emission, research in academia towards the time-gated Raman spectrometer, and fluorescence suppressed

Raman spectra, have been successfully observed for both Kerr gate and ICCDs. As discussed in section 2.3.2, the CMOS SPAD is a promising detector for its single photon sensitivity and integration capability with CMOS control circuits. Operated in the time-gated mode, the SPAD can be controlled to perform detection only in a predetermined short time window and reject unwanted photons arriving out of the time window, as shown in Figure 2.12. Owing to its low-cost, high sensitivity and fast gating capability, a time-gated (TG) CMOS SPAD is designed for the proposed system. To perform time-gated detection, a pulsed laser will be used.

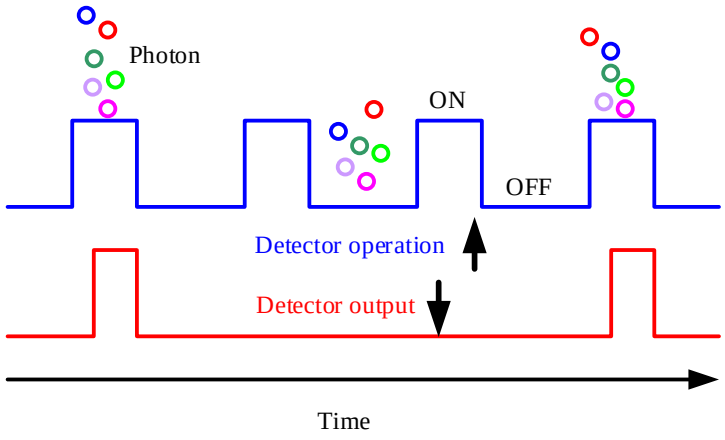


Figure 2.12: Time-gated operation of SPAD

2.4.2 System Operation and Specifications

With the main system components determined, the system is designed, as shown in Figure 2.13. The incident light from a pulsed laser passes through a dichroic mirror and is focused onto the sample by the sample illumination optics. The fluorescence, Stokes, anti-Stokes and Rayleigh scattered signals are collected by the light collection optics and then separated by the dichroic mirror, with the shorter wavelength components (anti-Stokes Raman and Rayleigh) transmitted and the longer wavelength components (Stokes scattering and fluorescence) reflected. The reflected signal is then coupled into an optical fiber by the second lens and forwarded to the Rowland circle of the concave grating. Finally, the CMOS TG-SPAD is positioned on the Rowland circle to acquire the Raman spectrum.

According to Figure 2.13, design considerations of this TG Raman spectrometer include the excitation source, concave grating and detector. Important parameters of the excitation source are the wavelength, pulse width, and repetition rate. A visible-light laser will be used for purpose of improving the Raman scattering efficiency and matching the wavelength response of the detector.

For CW Raman spectrometers, a variety of visible sources have been used, such as 488nm [112], 514nm [86], 532nm [113] and 633nm [114].

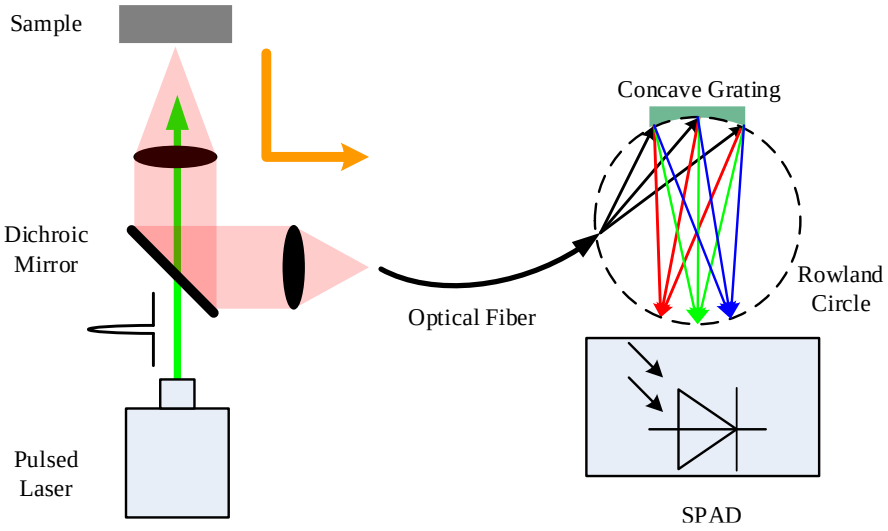


Figure 2.13: Optical diagram of the proposed TG Raman spectrometer

With regard to the time-gated Raman spectrometer, the pulsed green laser (532nm) has been widely used, as shown in Table 2.4. Both pulse width and repetition rate are very important for the suppression of fluorescence signals. In principle, the pulse width is usually selected to be shorter than the rising edge of the fluorescence emission (Figure 2.9), typically in the range of hundreds of picoseconds. The maximum repetition rate is determined by the fluorescence lifetime. However, limited by the maximum operation frequency of the laser source, usually a low repetition rate is used.

The Raman shifts of most chemicals are between 500cm^{-1} and 3000cm^{-1}, which corresponds to a wavelength band from 546nm to 633nm if a green laser (532nm) is used. The concave grating can be designed according to this wavelength band. Design considerations of a concave grating include the grating constant, radius, and other system parameters, such as the incident angle, entrance slit width and numerical aperture. The optimum combination of these parameters can be determined through simulations of the diffraction efficiency and spectral resolution. Since this work is targeting a low-cost system, the overall performance of the concave grating is also limited by the available fabrication process.

The parameter dominating the application of a SPAD is the wavelength response. As mentioned above, the main reason to choose a visible excitation is to improve the Raman scattering efficiency and match the wavelength response of Si SPADs. With reference to Figure 2.5(a), the target

wavelength band (546-633nm) is covered by the high QE region for different depletion widths. In addition, other parameters important to the SPAD are the photon detection efficiency (PDE), dark count rate (DCR), deadtime, and fill factor. Deadtime is related to the gate window and reset time of the SPAD, which can be controlled. However, there is not enough freedom to control the PDE and DCR of SPAD in standard CMOS technology, detailed information about these parameters will be given in Chapter 4. Table 2.6 summarizes the specifications of the major components of the proposed system. A pulsed laser with hundreds picoseconds pulse width is selected, narrower than the rising edge of fluorescence emission. The repetition rate is determined by both the fluorescence lifetime of the sample and the maximum repetition rate of the laser. With the excitation wavelength and Raman shift, the wavelength selector can be designed on basis of the wavelength band. Specifications of the wavelength selector will be given in chapter 3.

Table 2.6 Target specifications of proposed system

Component	Parameter	Specification
Excitation source	Wavelength	532nm
	Pulse width	~300-500ps
	Repetition rate	~kHz
Concave grating	Grating radius	~cm
	Grating constant	~800nm
	Wavelength band	546-633nm
TG-SPAD	Gating frequency	Max @ 100MHz
	Gate window	Max @ ns
	Pixel active region	10 x 10μm^2

2.5 Summary

In this chapter, the basic theory of Raman spectroscopy and a review of state-of-the-art Raman instruments and advanced Raman techniques were introduced. Derived from the knowledge for Raman scattering and instruments, a compact and low-cost Raman spectrometer was proposed by combining a concave grating and a CMOS SPAD. To suppress the background fluorescence without using a NIR excitation source, the CMOS SPAD will be operated in the time-gated mode to suppress fluorescence background and acquire Raman signal when a sample is excited by a very short laser pulse. The following chapter will introduce the implementation of the concave grating, its design and fabrication.

Chapter 3

Wavelength Selector

The function of a wavelength selector is to provide spatial light wavelength separation according to the wavelength. Wavelength separation can be achieved by a diffraction grating or a prism, so that different wavelengths propagate in different directions after the selector. To minimize the system, a concave diffraction grating is designed and fabricated.

3.1 Theoretical Background of Diffraction Grating

A diffraction grating is a periodic structure which consists of a large number of grooves with micrometer or sub-micrometer dimensions. When a light beam is incident on the surface of the grating, each groove can be taken as a small "point" source of reflected/transmitted light depending on the type of grating. The reflected/transmitted light from successive grooves combines and interferes to form set of diffracted wavefronts. According to grating theory, for a specific groove distance d and incident angle α, there exists a set of discrete angles, along which the reflected/transmitted light from all grooves are in phase, forming constructive interference patterns [115].

Figure 3.1 shows the diffraction diagram of a reflective planar grating. A light source with wavelength λ is incident on the grating with incident angle α to the grating normal. Based on the grating constant d and the wavelength λ, light is diffracted to different directions according to the well-known grating equation,

$$m\lambda = d\left(\sin\alpha + \sin\beta_m\right). \tag{3.1}$$

The right-hand term in Eq. (3.1) refers to the optical path difference between light from adjacent grooves. Constructive interference only occurs when this difference is equal to the multiples m of the incident wavelength, which corresponds to different diffraction angles β_m and diffraction order

m, as shown in Figure 3.1. According to Eq. (3.1), the diffraction angle β_m can be calculated by Eq. (3.2), where $G=1/d$ is defined as the groove density (grooves/mm).

$$\beta_m = \sin^{-1}\left(mG\lambda - \sin\alpha\right) \tag{3.2}$$

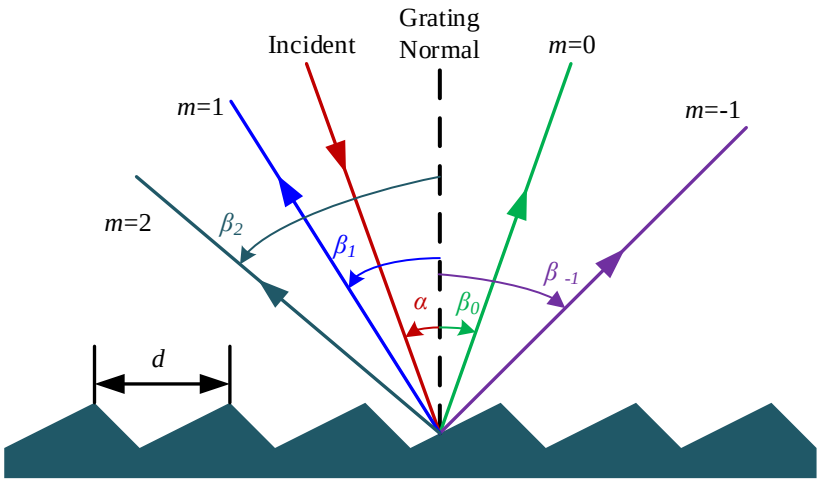


Figure 3.1: Diffraction of a reflective grating

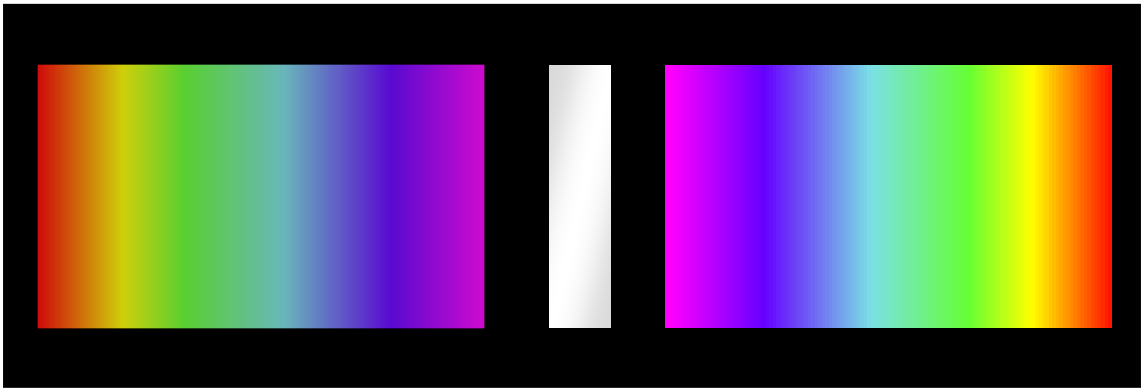


Figure 3.2: Spectrum of a white source dispersed by a diffraction grating

For a specific diffraction order m, the diffraction angle β_m is wavelength dependent. Consequently, light with different wavelengths can be spatially separated along the diffraction angles, and this is known as the dispersion property of a diffraction grating. However, there exists a unique diffraction order m=0, diffraction angle of the 0th order is wavelength independent ($\beta_0 = -\alpha$). Since all wavelength components overlap at the 0th order position, no dispersion occurs, and the 0th order is actually the reflection of the incident beam by the grating plane. If the incident light is white, the diffraction light of the 0th order will also be white, since it contains all wavelength components. Figure 3.2 shows the spectrum of a white source dispersed by a diffraction grating, from which we can see the different diffraction orders (rainbow) and the 0th order reflection (white).

3.1.1 Dispersion

The dispersion property defines the capability of a diffraction grating to spatially separate the wavelengths. Different parameters have been used to evaluate the dispersion property of a grating, such as the angular dispersion and linear dispersion. Angular dispersion measures the angular separation between successive wavelengths, and linear dispersion is the product of the angular dispersion and the effective focal length. For a given diffraction order and incident angle, the angular dispersion can be calculated by differentiating the grating equation (Eq. (3.1)),

$$\text{Angular Dispersion} = \frac{\mathrm{d}\beta}{\mathrm{d}\lambda} = \frac{mG}{\cos\beta}. \tag{3.3}$$

Substituting Eq. (3.2) in Eq. (3.3), the angular dispersion can be rewritten as

$$\text{Angular Dispersion} = \frac{\mathrm{d}\beta}{\mathrm{d}\lambda} = \frac{mG}{\cos\left(\sin^{-1}\left(mG\lambda - \sin\alpha\right)\right)}. \tag{3.4}$$

According to Eq. (3.4), the angular dispersion is a function of the groove density G. Broad angular spread can be achieved by increasing the groove density, with the constraint $|mG\lambda - \sin\alpha| < 1$.

3.1.2 Free Spectral Range

As introduced above, when a light source is incident on a grating surface, multi-order diffraction may occur depending on the incident wavelength $\lambda_{i\text{-}j}$ and the grating constant d. In case multi-order diffraction exists, it is possible for wavelengths from neighboring diffraction orders to be diffracted to the same diffraction angle. From the spectrum point of view, partial spectral overlapping occurs between adjacent diffraction orders. The wavelength range for which there is no overlapping is called the free spectral range.

Figure 3.3 demonstrates the spectral overlapping between adjacent diffraction orders. When designing a grating, the free spectral range should be appropriately evaluated, which can be calculated by using the grating equation Eq. (3.1). For a given diffraction order m and grating constant G, if λ_1 from the m^{th} order overlaps with λ_2 from the $(m+1)^{\text{th}}$ order, combining the grating equations of the two wavelengths gives

$$mG\lambda_1 = (m+1)G\lambda_2. \tag{3.5}$$

From Eq. (3.5), the free spectral range can be calculated by

$$\text{Free spectral range} = \Delta\lambda = \lambda_1 - \lambda_2 = \frac{\lambda_2}{m}. \tag{3.6}$$

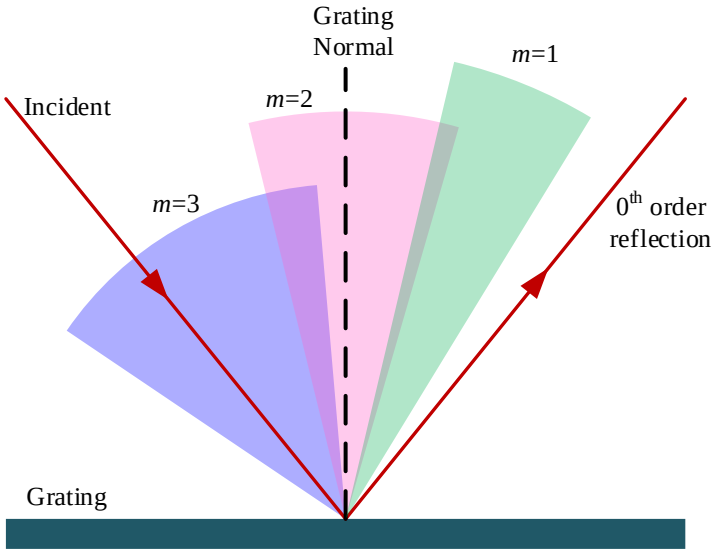


Figure 3.3: Diagram demonstrating the spectral overlapping of a grating

With Eq. (3.6), if the target wavelength is longer than 500nm, and the 1st diffraction order is measured, then the maximum wavelength that can be measured without overlapping is 2 x λ_2 = 1000nm. In application when the incident wavelength band is wider than the free spectral range, the superposition of wavelengths can affect the accuracy of the measured spectrum. Therefore, removing the overlapping wavelength band is important for accurate measurement of the spectrum. A common way to deal with the spectral overlapping is to use a filter to attenuate the wavelengths out of the free spectral range.

3.1.3 Diffraction Limit

As the most significant parameter of a wavelength selector, the spectral resolution is a function of a variety of parameters including the entrance slit, incident angle, grating, and exit slit. Among them, contribution from the grating can be characterized by the resolving power *P*, which is defined as the ability of a grating to separate adjacent wavelengths. If the minimum wavelength that can be distinguished at wavelength λ is $\Delta\lambda$, then the resolving power at this wavelength is [115]

$$P = \frac{\lambda}{\Delta\lambda} = |m| Gw, \tag{3.7}$$

where *w* is the width of the incident beam, and *Gw* refers to the number of grooves illuminated by the incident light. The contribution of the grating to the overall spectral resolution, which is also named diffraction limit, can be derived from Eq. (3.7) and is

$$\Delta\lambda_{Diff\_Lim} = \frac{\lambda}{|m|Gw}. \tag{3.8}$$

From Eq. (3.8), the resolution $\Delta\lambda_{Diff\_Lim}$ caused by the diffraction limit is dependent on both the diffraction order m and the groove density G. However, m and G are not independent, and as long as the grating equation is satisfied, better spectral resolution can be obtained with higher diffraction orders and groove density.

3.1.4 Diffraction Efficiency

When light is incident on a diffraction grating, it can be dispersed to different diffraction orders m, with different light intensities. The diffraction efficiency of a grating is defined as the ratio of the intensity I_m of the m^{th} order diffraction beam to the incident light intensity I_o. The classical way to describe diffraction efficiency is based on the Fresnel-Kirchhoff approximation, from which the diffraction efficiency (I_m/I_o) of a plane grating can be written as [116]

$$\frac{I_m}{I_0} = \left(\frac{\sin\sigma}{\sigma}\right)^2 \cdot \left(\frac{\sin N\frac{\delta}{2}}{\sin\frac{\delta}{2}}\right)^2, \tag{3.9}$$

$$\text{where } \sigma = \frac{\pi a(\sin\alpha_i + \sin\beta_i)}{\lambda},\ \delta = \frac{2\pi d(\sin\alpha + \sin\beta)}{\lambda}. \tag{3.10}$$

The diffraction efficiency in Eq. (3.9) is a product of two terms. The first term refers to single groove diffraction which governs the intensity. The second term is the interference between multiple grooves. N (N=Gw) in Eq. (3.9) is the number of grooves, a and d in Eq. (3.10) are the groove size and groove distance, α_i and β_i are the incident and diffraction angles to the groove normal, and α and β are the incident and diffraction angles to the grating normal. The single groove intensity term determines the energy distribution among different diffraction angles, and it is relevant to the shape of the groove. For gratings with rectangular grooves, the maximum intensity coincides with the 0^{th} diffraction order. For blazed gratings with triangular grooves, the maximum intensity position can be shifted to other diffraction orders depending on the angle of the grooves. In this design of the concave grating, characterization of the diffraction efficiency is performed with a commercial simulator.

3.2 Concave Grating Design

When designing a concave grating, two important parameters are the diffraction efficiency and the overall system's spectral resolution. As introduced in section 3.1, both the diffraction efficiency and spectral resolution depend on several parameters, including parameters of the system and the grating. Thus, the final goal of this design is to find an optimum combination of the design parameters for target specifications of diffraction efficiency, spectral resolution, and free spectral range. It will start with the characterization of the diffraction efficiency.

3.2.1 Diffraction Efficiency

According to equation Eq. (3.1), for a fixed groove density, a grating can perform the dispersion function only for a specific wavelength band and for certain diffraction orders, for which $|mG\lambda - \sin\alpha|<1$. In order words, when the wavelength band is fixed, only gratings with certain groove densities can be used. As given in section 2.4.2, for excitation with a green laser (532nm), the wavelength of the Raman signal is between 546nm and 633nm. The optimum grating constant for this wavelength band was determined through simulation of the diffraction efficiency.

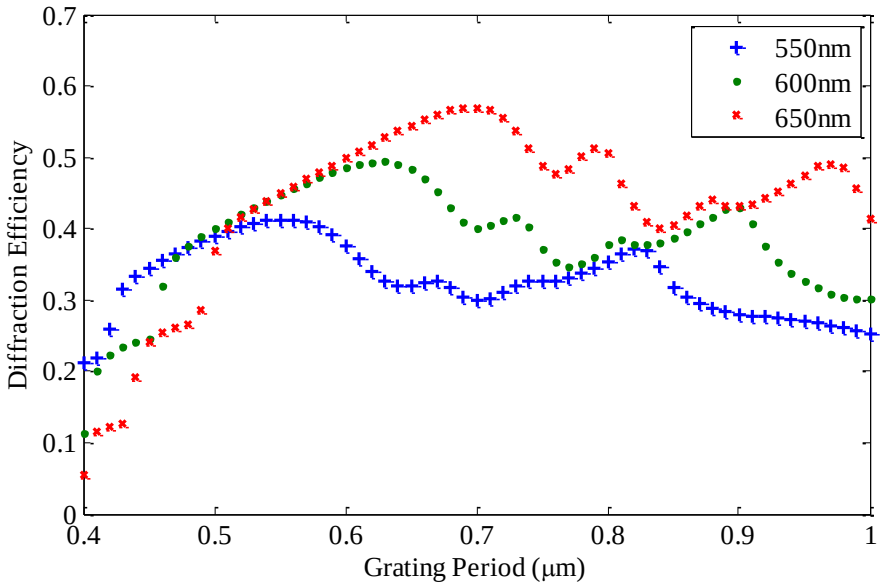


Figure 3.4: Simulation of the diffraction efficiency of concave gratings with different grating periods, at other conditions: 30deg incident angle of light, Au coated, 25mm radius

Simulation of the diffraction efficiency was carried out with the commercial simulator PCGrate [117]. Figure 3.4 shows the diffraction efficiency *vs.* the grating period d from d=400nm to 1μm. Since the wavelength band of the proposed system is between 546nm to 633nm, three wavelengths

were simulated (550, 600 and 650nm). Taking into account the technology details of the fabrication process, a sinusoidal groove shape was selected for simulation.

In Figure 3.4, for a given wavelength, the diffraction efficiency varies with the grating period, peaking for particular values of the grating period. Increasing the wavelength, the peak diffraction efficiency shifts to larger grating periods. Because diffraction is enhanced when the groove size approaching the incident wavelength. Inspecting the overall response of the concave grating in the entire wavelength band, grating periods between 550nm and 1000nm were found to be suitable. Therefore, for the rest of the following simulations, a fixed grating period of 700nm was used.

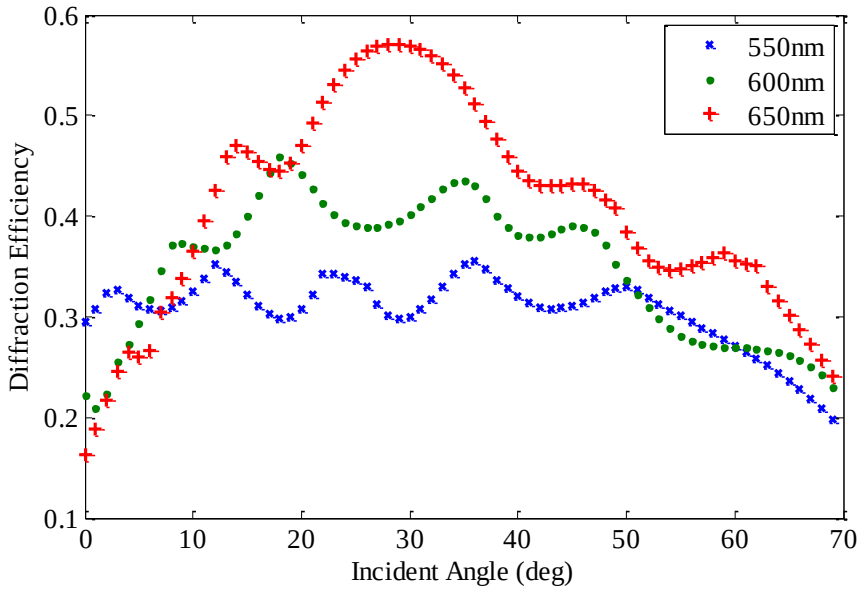


Figure 3.5: Simulation of the diffraction efficiency of a concave grating for different incident angles, at other conditions: 700nm grating period, Au coated, 25mm radius

With the fixed grating period, further simulations were carried out to determine the dependence of the diffraction efficiency on the incident angle. Figure 3.5 shows the simulation results for the three selected wavelengths. The diffraction efficiency fluctuates with the incident angles from 0° to 70°. This fluctuation is due to the dependence of the reflection coefficient on the incident angle. Based on the simulation results, incident angles between 20° and 40° are preferred for higher diffraction efficiency.

A diffraction grating can be of either reflection or transmission type. In this work, a reflective grating was designed. To obtain a reflective surface, the grating is usually coated with a thin layer of reflective metal. There are many choices of metals for grating coating, such as silver (Ag), gold

(Au), or copper (Cu). The metal can be selected according to its optical properties, such as the refractive index.

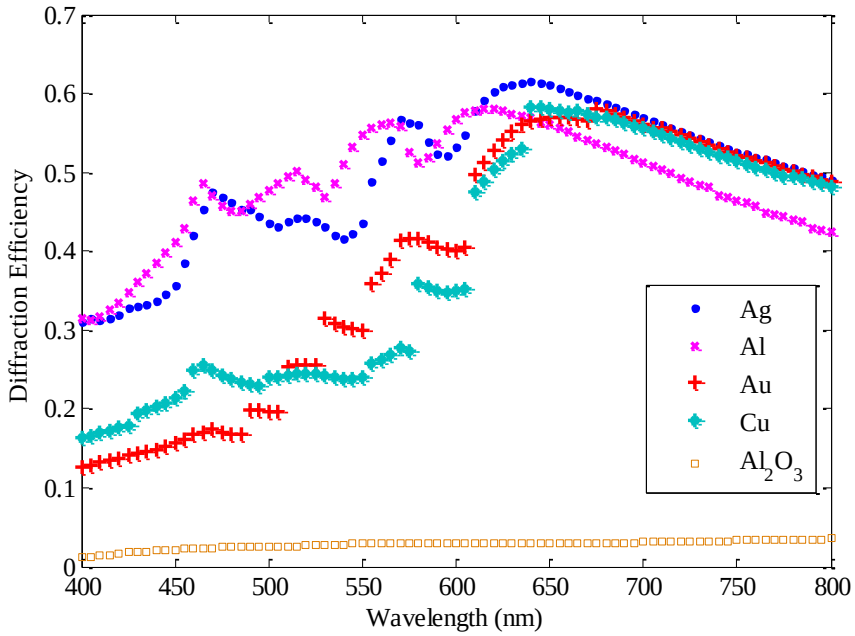


Figure 3.6: Simulation of the wavelength response of gratings with surfaces coated by different materials (700nm grating period, 25mm radius, 30deg incident angle)

Figure 3.6 shows the diffraction efficiency of grating coated by different materials. For short wavelengths in the visible region, Ag and Al coated gratings have higher diffraction efficiencies than gratings coated with Au or Cu. However, this difference diminishes when the wavelength increases to the NIR region. If only the diffraction efficiency is considered, then Ag or Al should be a better choice than Au. However, both Ag and Al easily oxidize, and Au is a most common choice for gratings, because of its chemical stability. After oxidation, the diffraction efficiency drops very fast to a very low value, as illustrated for Al_2O_3 in Figure 3.6. Therefore, Au was selected as the coating material of the concave grating.

3.2.2 Spectral Resolution

For a wavelength selection system, the spectral resolution is the minimum width $\Delta\lambda$ of the spectral line that can be resolved. The spectral resolution is determined by both the grating structural information (grating period and radius) and other system parameters, such as the incident angle, entrance slit width, and numerical aperture.

Figure 3.7 shows a concave grating-based wavelength selector, mounted on the Rowland circle. As mentioned in chapter 2, for the Rowland configuration, when the entrance slit is positioned on

the Rowland circle, light diffracted from the concave grating is also focused on the Rowland circle, but at different positions depending on the wavelengths. The main components in this Rowland mount are the concave grating, entrance slit (fiber) and exit slit (detector). For an inexpensive setup, the incident light is delivered directly from a fiber and the fiber diameter determines the width of the entrance slit, while the size of the detector determines the width of the exit slit. Thus, both slits are predetermined to be in the range of 10μm to 100μm. The numerical aperture of the fiber determines the numerical aperture of the system.

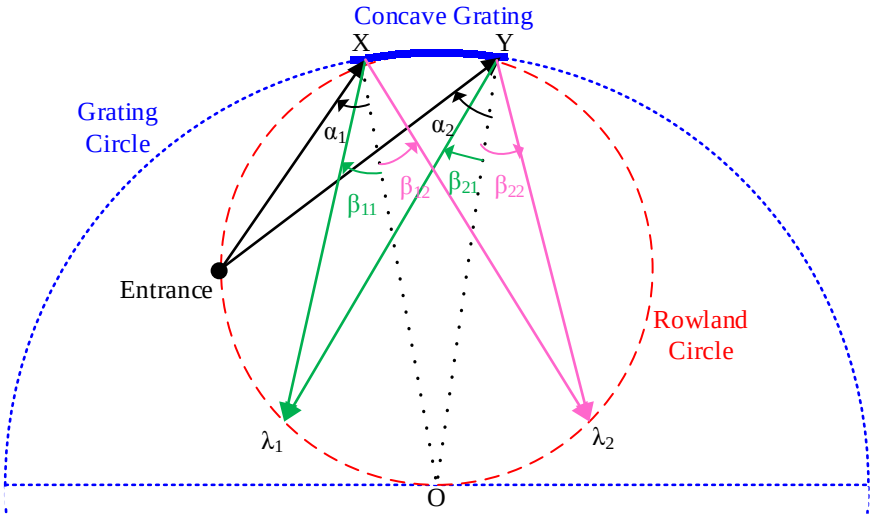


Figure 3.7: Rowland configuration of the concave grating system

According to [118], there are several contributions $\Delta\lambda_i$ in the total spectral resolution $\Delta\lambda$ of a grating-based system. These contributions include entrance slit width, exit slit width, aberration and diffraction limit. The square root of the quadratic sum of them gives the overall spectral resolution of the system. Since the diffraction beam from the concave grating is delivered by a fiber to the detector, the contribution from the exit slit was not considered. The contribution from the grating is the diffraction limit of the grating, which is $\Delta\lambda_{Diff\_Lim}$, given by Eq. (3.8) earlier. Resolutions caused by entrance slit and aberration can be calculated [118] from

$$\Delta\lambda_{Entrance} = \frac{Sd\cos\alpha}{mr_{in}}, \tag{3.11}$$

$$\text{and } \Delta\lambda_{Aberration} = \frac{\Delta x d\cos\beta}{mr_b}. \tag{3.12}$$

In Eq. (3.11), S is the entrance slit width, and r_{in} is the distance from entrance slit to grating center. In Eq. (3.12), Δx is the horizontal deviation of the image caused by aberration, and r_b is the distance from grating center to the detection plane.

Regarding the calculations of the different contributions $\Delta\lambda_i$, both $\Delta\lambda_{Entrance}$ and $\Delta\lambda_{Diff\_Lim}$ can be calculated when the grating and grating mount are fixed. To investigate the total spectral resolution, the parameter unknown and to be solved is the aberration induced horizontal deviation Δx. Thus, calculation of the aberration contribution is of great importance. Before calculating the spectrum broadening due to aberration, aberration theory is discussed.

3.2.2.1 Aberration

Aberration in a Rowland mount is related to the focusing property of the concave grating. As depicted in Figure 3.7, points X and Y are two arbitrary points on the grating curvature and OX and OY are the normal vectors at these points. If the arc length XY is small enough compared to the grating radius R, then X and Y can be assumed to be on the Rowland circle. In this case, the incident angles α_1 and α_2 are identical.

According to the grating Eq. (3.1), for a certain diffraction order m, wavelength λ and groove density G, the same incident angles have the same diffraction angles, which means that β_{11} and β_{21} in Figure 3.7 are also identical. Therefore, light of wavelength λ diffracted from different points of the grating will intersect with the Rowland circle at the same point, which gives rise to the focusing property of a concave grating. However, since X and Y are not on the Rowland circle, this approximation will caused the light diffracted from the grating to intersect with the Rowland circle at slightly different positions.

Figure 3.8 shows the geometry of a concave reflection grating with a point source A (x_A y_A 0), where x-axis is perpendicular and z-axis is parallel to the grooves. Light diffracted from point O (0 R 0) and point P (x_p y_p z_p) intersects with the detection plane at B_0 and B, respectively. Ideally, if there is no aberration, B and B_0 will overlap. However, an aberration free image is rare and there is deviation between B and B_0. This deviation is caused by aberration of the concave grating. Aberration in a concave grating can be calculated by solving the optical path difference (OPD) function, which can be written as

$$OPD = F = APB - AOB_0 + mN\lambda, \tag{3.13}$$

where N is the number of grooves between point O and point P. According to Eq. (3.13), zero aberration means OPD is equal to zero, so better image can be achieved by minimizing OPD. The OPD can be analyzed through the Tylor series expansion in terms of x_p and z_p [Eq. (3.14)], assuming a constant line space grating.

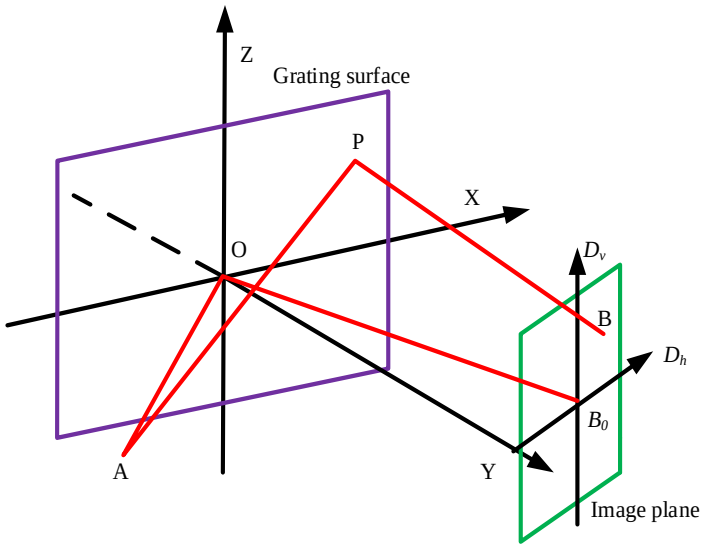


Figure 3.8: Geometry of a concave grating: A-point source, O-grating center, B-image of point source in image plane, P-arbitrary point on grating surface

$$OPD = F = APB - AOB_0 + mN\lambda, \tag{3.13}$$

where N is the number of grooves between point O and point P. According to Eq. (3.13), zero aberration means OPD is equal to zero, so better image can be achieved by minimizing OPD. The OPD can be analyzed through the Tylor series expansion in terms of x_p and z_p [Eq. (3.14)], assuming a constant line space grating.

$$F = \sum_{i,j} \frac{x_p^i z_p^i}{i!\,j!} F_{ij}, \tag{3.14}$$

where $F_{ij} = \dfrac{\partial^{i+j}(APB - AOB_0)}{\partial x^i \partial z^j}$, and $\dfrac{\partial^i (mN\lambda)}{\partial x^i} = \dfrac{\partial^j (mN\lambda)}{\partial z^j} = 0$ for $(i \geq 2)$.

Each term in the series is related to a different type of aberration. If the series coefficient for one F_{ij} is equal to zero, then the aberration associated with that term will disappear. Eq. (3.15) gives some of the low order terms with which the horizontal (D_h) and vertical (D_v) deviations can be calculated by Eq. (3.16).

$$F_{10} = -\sin\alpha - \sin\beta + m\lambda G,\ F_{01} = \frac{z_A}{r_{in}} + \frac{z_b}{r_b}$$

$$F_{20} = \left(\frac{\cos^2\alpha}{r_{in}} - \frac{\cos\alpha}{R}\right) + \left(\frac{\cos^2\beta}{r_b} - \frac{\cos\beta}{R}\right),\ F_{02} = \left(\frac{1}{r_{in}} - \frac{\cos\alpha}{R}\right) + \left(\frac{1}{r_b} - \frac{\cos\beta}{R}\right)$$

$$F_{30} = 3\frac{\sin\alpha}{r_{in}}\left(\frac{\cos^2\alpha}{r_{in}} - \frac{\cos\alpha}{R}\right) + 3\frac{\sin\beta}{r_b}\left(\frac{\cos^2\beta}{r_b} - \frac{\cos\beta}{R}\right)$$

$$F_{12} = \frac{\sin\alpha}{r_{in}}\left(\frac{1}{r_{in}} - \frac{\cos\alpha}{R}\right) + \frac{\sin\beta}{r_b}\left(\frac{1}{r_b} - \frac{\cos\beta}{R}\right)$$

$$F_{04} = \frac{3\sin^2\alpha}{R^2 r_{in}} - \frac{3\cos\alpha}{R^3} - \frac{3}{{r_{in}}^3} + \frac{6\cos\alpha}{R{r_{in}}^2} + \frac{3\sin^2\beta}{R^2 r_b} - \frac{3\cos\beta}{R^3} - \frac{3}{{r_b}^3} + \frac{6\cos\beta}{R{r_b}^2}$$

$$F_{40} = \frac{3\cos^2\alpha}{{r_{in}}^3}\left(5\sin^2\alpha - 1\right) + \frac{6\cos\alpha}{R{r_{in}}^2}\left(1 - 3\sin^2\alpha\right) + \frac{3\sin^2\alpha}{R^2 r_{in}} - \frac{3\cos\alpha}{R^3} \ldots$$
$$\ldots + \frac{3\cos^2\beta}{{r_b}^3}\left(5\sin^2\beta - 1\right) + \frac{6\cos\beta}{R{r_b}^2}\left(1 - 3\sin^2\beta\right) + \frac{3\sin^2\beta}{R^2 r_b} - \frac{3\cos\beta}{R^3}$$

$$F_{22} = \frac{\sin^2\alpha}{R^2 r_{in}} + \frac{3\sin^2\alpha - 1}{{r_{in}}^3} + \frac{\cos\alpha\left(2 - 3\sin^2\alpha\right)}{R{r_{in}}^2} - \frac{\cos\alpha}{R^3} + \frac{\sin^2\beta}{R^2 r_b} + \frac{3\sin^2\beta - 1}{{r_b}^3} + \frac{\cos\beta\left(2 - 3\sin^2\beta\right)}{R{r_b}^2} - \frac{\cos\beta}{R^3}$$

$$F_{50} = \frac{15\sin\alpha}{R^2{r_{in}}^2}\left(1 - 3\cos^2\alpha\right) + \frac{15\sin\alpha}{{r_{in}}^4}\left(10\sin^2\alpha - 3 - 7\sin^4\alpha\right) - \frac{15\sin\alpha\cos\alpha}{R^3 r_{in}} + \frac{30\sin\alpha\cos\alpha}{R{r_{in}}^3}\left(3 - 5\sin^2\alpha\right) \ldots$$
$$\ldots \frac{15\sin\beta}{R^2{r_b}^2}\left(1 - 3\sin^2\beta\right) + \frac{15\sin\beta}{{r_b}^4}\left(10\sin^2\beta - 3 - 7\sin^4\beta\right) - \frac{15\sin\beta\cos\beta}{R^3 r_b} + \frac{30\sin\beta\cos\beta}{R{r_b}^3}\left(3 - 5\sin^2\beta\right)$$

$$F_{32} = \frac{3\sin\alpha}{R^2{r_{in}}^2}\left(1 - 3\cos^2\alpha\right) + \frac{3\sin\alpha}{{r_{in}}^4}\left(5\sin^2\alpha - 3\right) - \frac{3\sin\alpha\cos\alpha}{R^3 r_{in}} + \frac{3\sin\alpha\cos\alpha}{R{r_{in}}^3}\left(6 - 5\sin^2\alpha\right) \ldots$$
$$\ldots \frac{3\sin\beta}{R^2{r_b}^2}\left(1 - 3\cos^2\beta\right) + \frac{3\sin\beta}{{r_b}^4}\left(5\sin^2\beta - 3\right) - \frac{3\sin\beta\cos\beta}{R^3 r_b} + \frac{3\sin\beta\cos\beta}{R{r_b}^3}\left(6 - 5\sin^2\beta\right)$$

$$F_{14} = \frac{3\sin\alpha}{R^2{r_{in}}^2}\left(1 - 3\cos^2\alpha\right) - \frac{9\sin\alpha}{{r_{in}}^4} + \frac{3\sin\alpha\cos\alpha}{Rr_{in}}\left(\frac{6}{{r_{in}}^2} - \frac{1}{R^2}\right) + \cdots$$
$$\cdots \frac{3\sin\beta}{R^2{r_b}^2}\left(1 - 3\cos^2\beta\right) - \frac{9\sin\beta}{{r_b}^4} + \frac{3\sin\beta\cos\beta}{Rr_b}\left(\frac{6}{{r_b}^2} - \frac{1}{R^2}\right) \tag{3.15}$$

The vertical and horizontal deviations can be calculated by

$$D_v = r_b \frac{\partial F}{\partial z_p} = r_b \left(z_p F_{02} + x_p z_p F_{12} + \frac{1}{6} z_p^3 F_{04} + \frac{1}{2} z_p x_p^2 F_{22} + \ldots \right)$$

$$D_h = \frac{r_b}{\cos\beta} \frac{\partial F}{\partial x_p} = \frac{r_b}{\cos\beta} \left(x_p F_{20} + \frac{1}{2} x_p^2 F_{30} + \frac{1}{2} z_p^2 F_{12} + \frac{1}{6} x_p^3 F_{40} + \frac{1}{2} x_p z_p^2 F_{22} + \ldots \right) \quad (3.16)$$

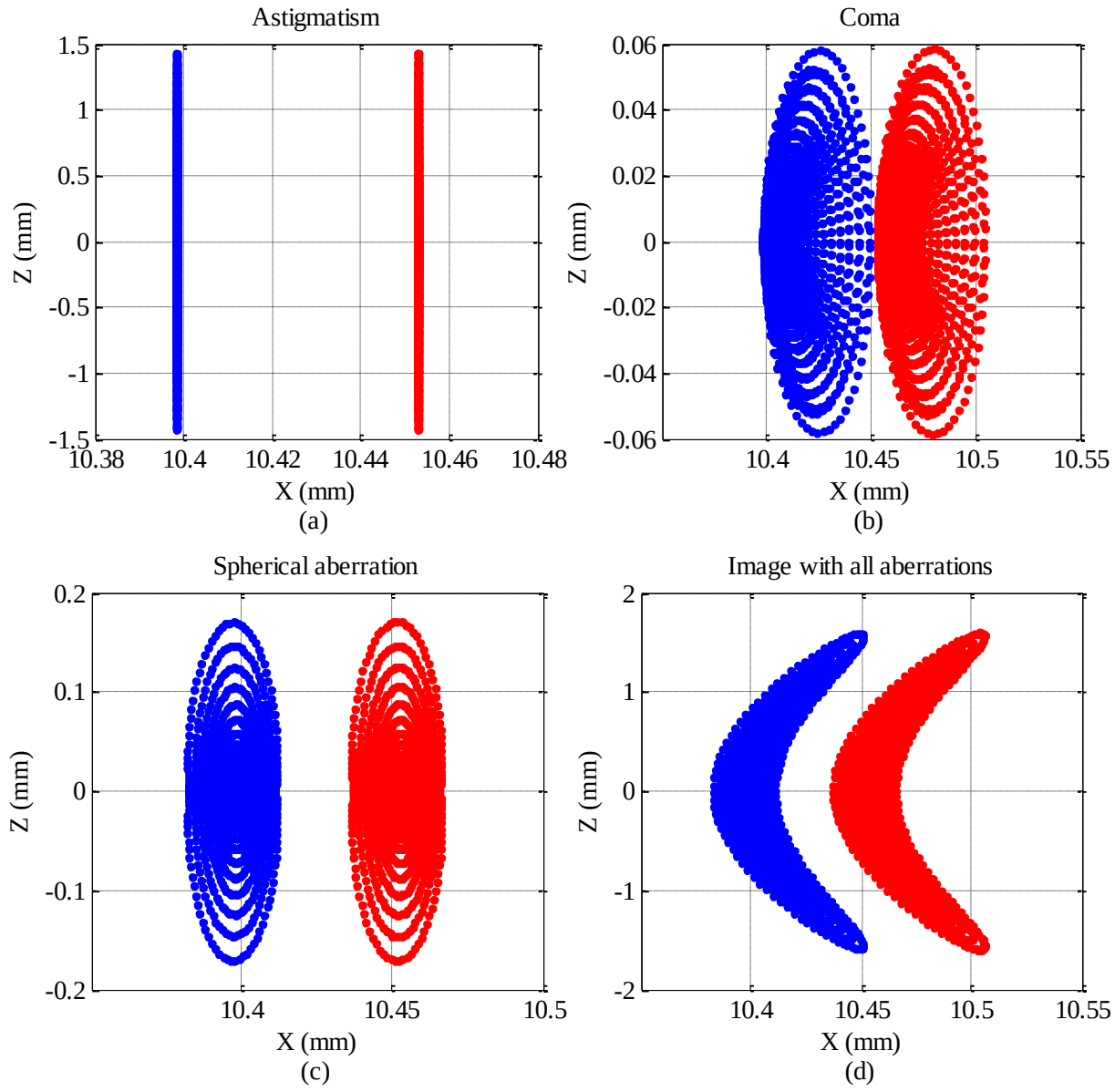


Figure 3.9: Spot diagrams of a point source with different aberration (λ=600, 602nm): (a) astigmatism; (b) coma; (c) spherical aberration; (d) image with all aberrations

There are 5 types of aberrations, line curvature (i+j=1), defocus (i=0, j=2), astigmatism (i=2, j=0), coma (i+j=3), spherical aberration (i+j=4) and other high-order aberrations. Each type of aberration can be characterized through Eqs. (3.14-3.16). Figure 3.9 shows the spot diagram of the

lower order aberrations and the spot diagram with all aberrations. Two wavelengths are simulated (600, 602nm), depicted with different colors in the figure. Other parameters included in this simulation are grating constant (700nm), incident angle (35deg), grating radius (25mm), fiber diameter (50μm), and numerical aperture (0.12). As shown in the simulation results, astigmatism can cause severe expanding of the image in the *z* direction, and the broadening of the image in *x* direction is mainly caused by coma. To obtain a high quality image, it is necessary to reduce the lower order aberrations when designing the system.

3.2.2.2 Spectral Resolution

As introduced above, there are basically three major sources for spectral broadening in the proposed system, including the contributions from diffraction limit ($\Delta\lambda_{Diff\_Lim}$), aberration ($\Delta\lambda_{Aberration}$), and entrance slit ($\Delta\lambda_{Entrance}$). The overall system spectral resolution is the square root of the quadratic sum of these three sources.

$$\Delta\lambda = \sqrt{\Delta\lambda_{Entrance}^2 + \Delta\lambda_{Aberration}^2 + \Delta\lambda_{Diff\_Lim}^2} \tag{3.17}$$

Among the three sources, the contribution from aberration is the most difficult to calculate. Since the Taylor series coefficient for F_{ij} given in Eq. (3.14) decreases with the increase of series orders ($i+j$), its contribution to the overall spectral broadening also decreases, so only lower order aberrations are considered in this design. Coefficients given in Eq. (3.15) are too complicated for calculation of the aberration contribution. Therefore, simplified algorithms will be proposed in this section.

Figure 3.10 shows the image of a point source considering aberration, with the same design parameters as in Figure 3.9. Assuming a point source with certain numerical aperture is used, then an illumination area with height *h* and width *w* on the concave grating can be obtained (see Figure 3.10(a)). As depicted in this figure, two situations exist when evaluating the spectral resolution. Normally, the height *h* is close to the width *w* or in the same order of magnitude. Light diffracted from A and B intersect with the image plane at A' and B', respectively. In this case, the largest horizontal deviation, which determines $\Delta\lambda_{Aberration}$, will be caused by A, rather than B. If z_p is the vertical coordinate of A, then the horizontal coordinate x_p will be very small. Therefore, Eq. (3.16) can be simplified as Eq. (3.18) and is given by

$$D_h = \frac{r_b}{\cos\beta}\frac{\partial F}{\partial x_p} = \frac{r_b}{\cos\beta}\left(\frac{1}{2}z_p^2 F_{12} + \frac{1}{24}z_p^4 F_{14} + \cdots\right). \tag{3.18}$$

Both F_{12} and F_{14} are already given in Eq. (3.15). If the higher order aberrations are neglected, then the aberration induced spectral resolution can be calculated by combining Eq. (3.12) and Eq. (3.15), to give

$$\Delta\lambda_{Aberration} = \frac{d}{m}\left(\frac{1}{2}z_p^2 F_{12} + \frac{1}{24}z_p^4 F_{14}\right). \tag{3.19}$$

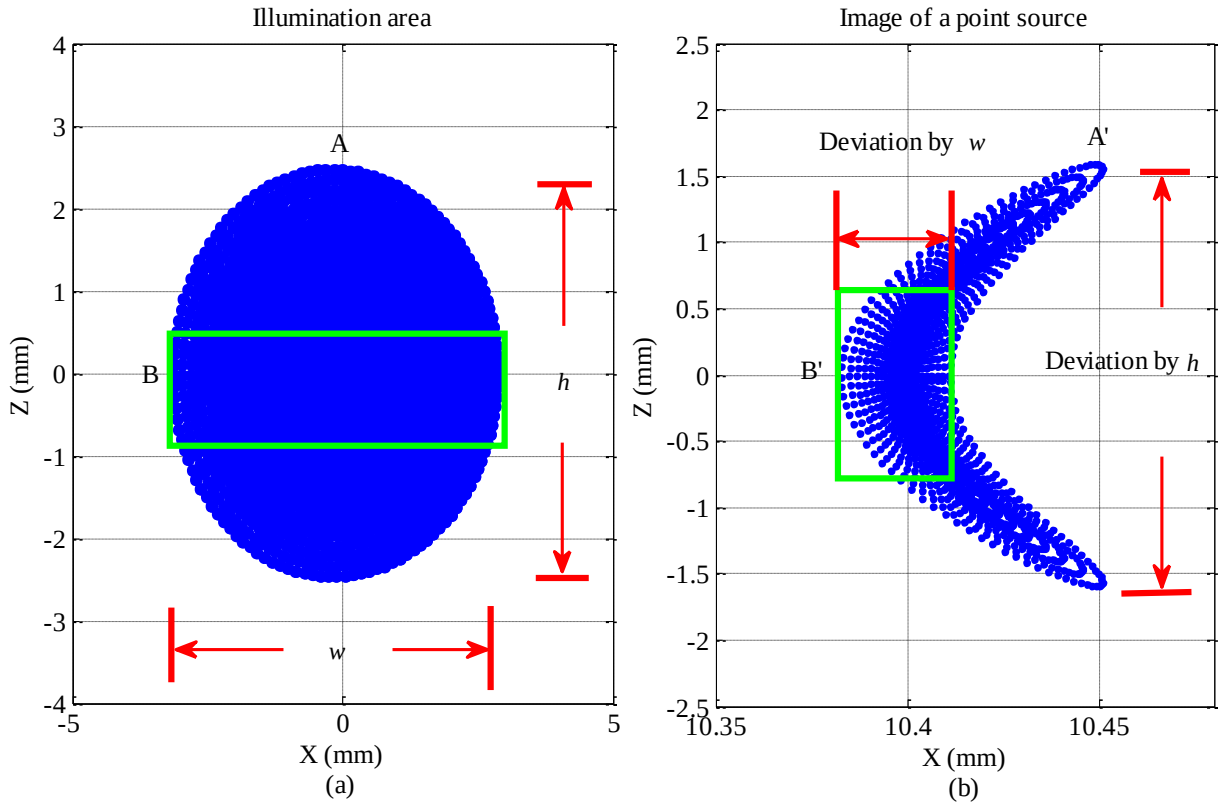


Figure 3.10: Spot diagram of: (a) Illuminated area of a point source on the concave grating surface; (b) Image of the corresponding point source

Since the horizontal deviation D_h in Eq. (3.18) is mainly caused by F_{12} and F_{14}, which play dominant role in spectrum broadening in horizontal direction. Figure 3.11 shows the simulation results of the variations of F_{12} and F_{14} as a function of incident angles and the aberration contribution ($\Delta\lambda_{Aberration}$) when considering F_{12} and (F_{12} + F_{14}), respectively.

As shown in Figure 3.11(a), F_{12} and F_{14} are with the same order of magnitude, but different sign. However, considering the fact that z_p is also very small, this causes the first term on right side of Eq. (3.19) to be larger than the second term. From simulations, there is negligible difference in the results with and without F_{14}, as shown in Figure 3.11(b). Therefore, only F_{12} is adequate to

quantify the aberration contribution. For the purposes of further simplifying the calculation process, F_{12} given in Eq. (3.15) is simplified, and Eq. (3.19) can now be rewritten as

$$\Delta\lambda_{Aberration} = \frac{1}{2}\frac{d}{m}z_p^2\left(\frac{\sin\alpha}{r_{in}}\left(\frac{1}{r_{in}}-\frac{\cos\alpha}{R}\right)+\frac{mG\lambda-\sin\alpha}{r_b^{\,2}}-\frac{mG\lambda-\sin\alpha}{Rr_b}+\frac{\left(mG\lambda-\sin\alpha\right)^3}{2Rr_b}\right). \quad (3.20)$$

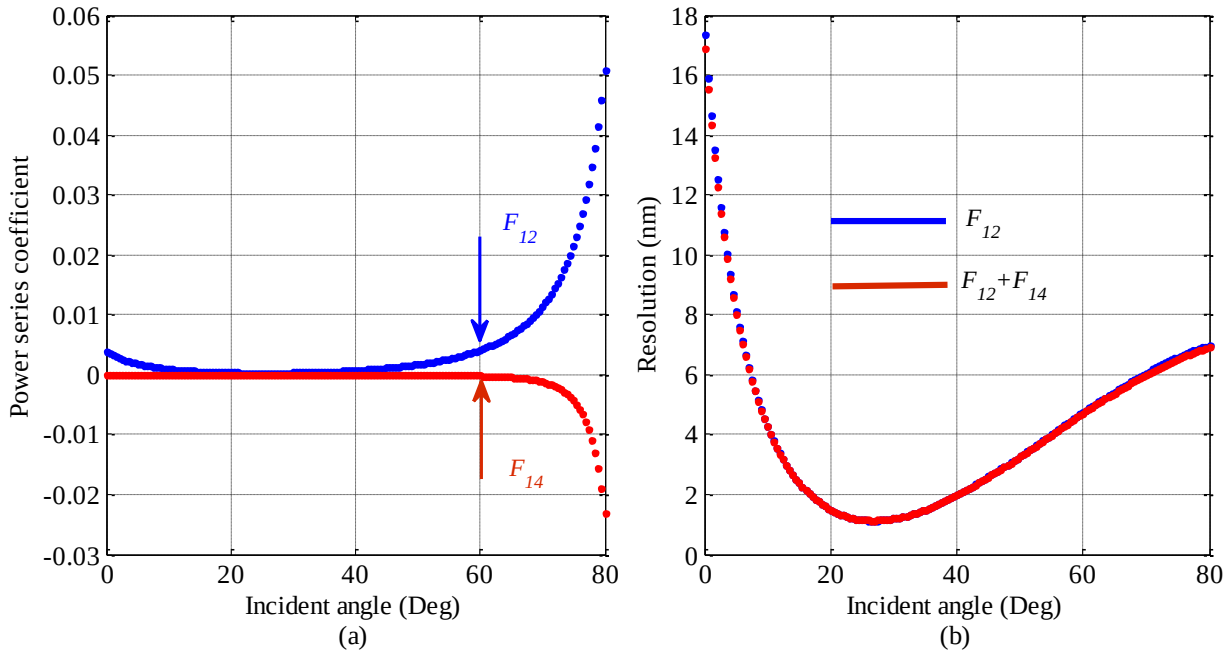


Figure 3.11: Situation 1 (25mm radius, 700nm grating constant, 0.12 NA): (a) Power series coefficients with incident angle. (b) Aberration related spectral resolution when considering different Taylor series coefficients.

In some situations, mirrors are used to restrict light transmission in some direction, functioning like a waveguide [52]. With mirrors, the spot height *h* will be much smaller than the spot width *w* (green area in Figure 3.10). Therefore, the largest horizontal deviation will be caused by point B, rather than point A. If (x_p y_p z_p) is the coordinate of point B, z_p will be close to zero. In this situation, the aberration relevant contribution can be calculated by using Eq. (3.21) (Rowland configuration, F_{20}=F_{30}=0), and is

$$\Delta\lambda_{Aberration} = \frac{d}{m}\frac{r_b}{\cos\beta}\frac{\partial F}{\partial x_p} = \frac{d}{m}\frac{r_b}{\cos\beta}\left(\frac{1}{6}x_p^3F_{40}+\frac{1}{24}x_p^4F_{50}+\frac{1}{120}x_p^5F_{60}+\cdots\right). \quad (3.21)$$

Considering F_{40}, F_{50}, and F_{60}, it is very complicated to calculate the aberration induced resolution. To evaluate the contribution of each of the three terms to the overall system spectral resolution, Figure 3.12 shows the simulation results of these coefficients as a function of incident angle and the associated contributions to the overall spectral resolution.

As shown in Figure 3.12(a), the three power series coefficients (F_{40}, F_{50}, and F_{60}) vary with the incident angle with the same trends. Figure 3.12(b) shows the variation of the aberration induced resolution with incident angle when considering F_{40}, (F_{40}+F_{50}) and (F_{40}+F_{50}+F_{60}). According to the calculation, deviation mainly occurs when using a large (>60°), or a very small incident angle (<10°). From the diffraction efficiency simulations, the incident angle of the proposed system is selected to be between 20° and 40°, in which case, only the F_{40} term is enough to calculate the aberration contribution. With the Rowland configuration, F_{40} in Eq. (3.14) can be further simplified as

$$F_{40} = \frac{3}{R^3}\left(\frac{1}{\cos\alpha} - \cos\alpha\right) + \frac{3\left(mG\lambda - \sin\alpha\right)^2}{R^3\left(1 - \frac{1}{2}\left(mG\lambda - \sin\alpha\right)\right)}. \tag{3.22}$$

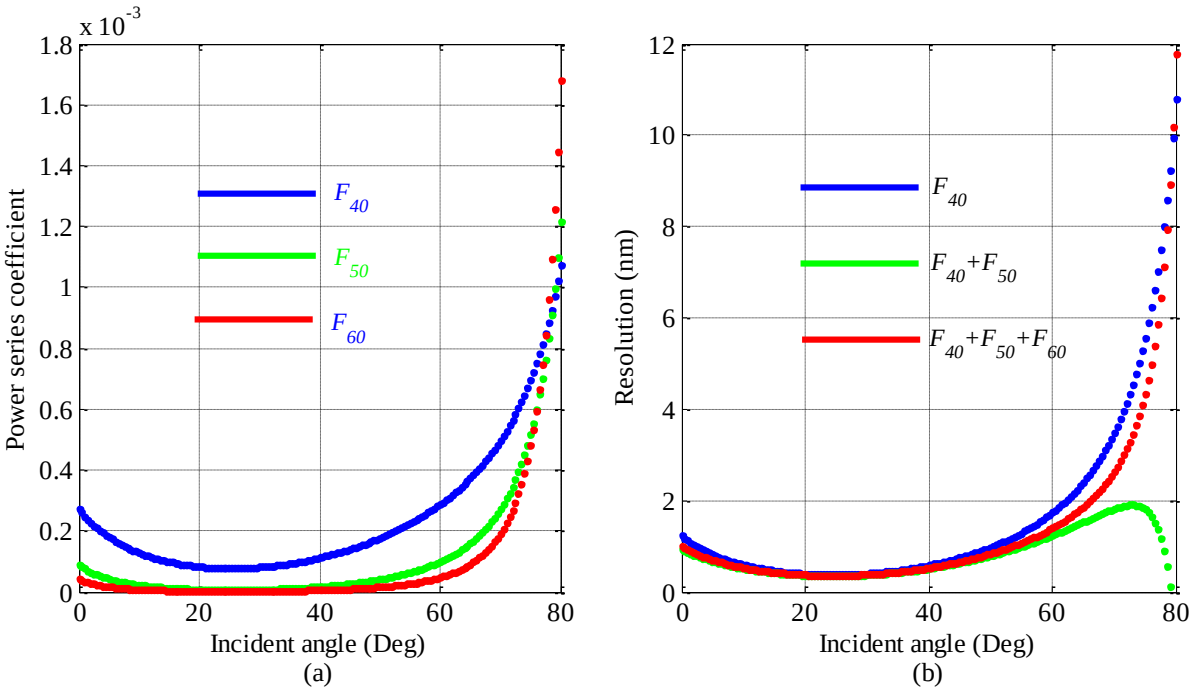


Figure 3.12: Situation 2 (25mm radius, 700nm grating constant, 0.12 NA): (a) Power series coefficients with incident angle. (b) Aberration related spectral resolution when considering different power series coefficients.

Considering only F_{40}, Eq. 3.21 has been tested using the parameters in [52]. The result of the calculation is ~1 nm spectral resolution, which is in very good agreement with both calculations (0.9 nm) and measurements (1.1 nm) reported in [52].

Combining the contributions from aberration, diffraction limitation and entrance slit, the overall spectral resolution can be obtained using Eq. (3.17). From this derivation, the overall spectral resolution is a function of several design parameters. Figure 3.13 shows the dependence

of the total spectral resolution on different system parameters of both situations discussed above – when the height *h* is close to the width *w* (Figure 3.13(a)), or when *h* is much smaller than *w* (Figure 3.13(b)).

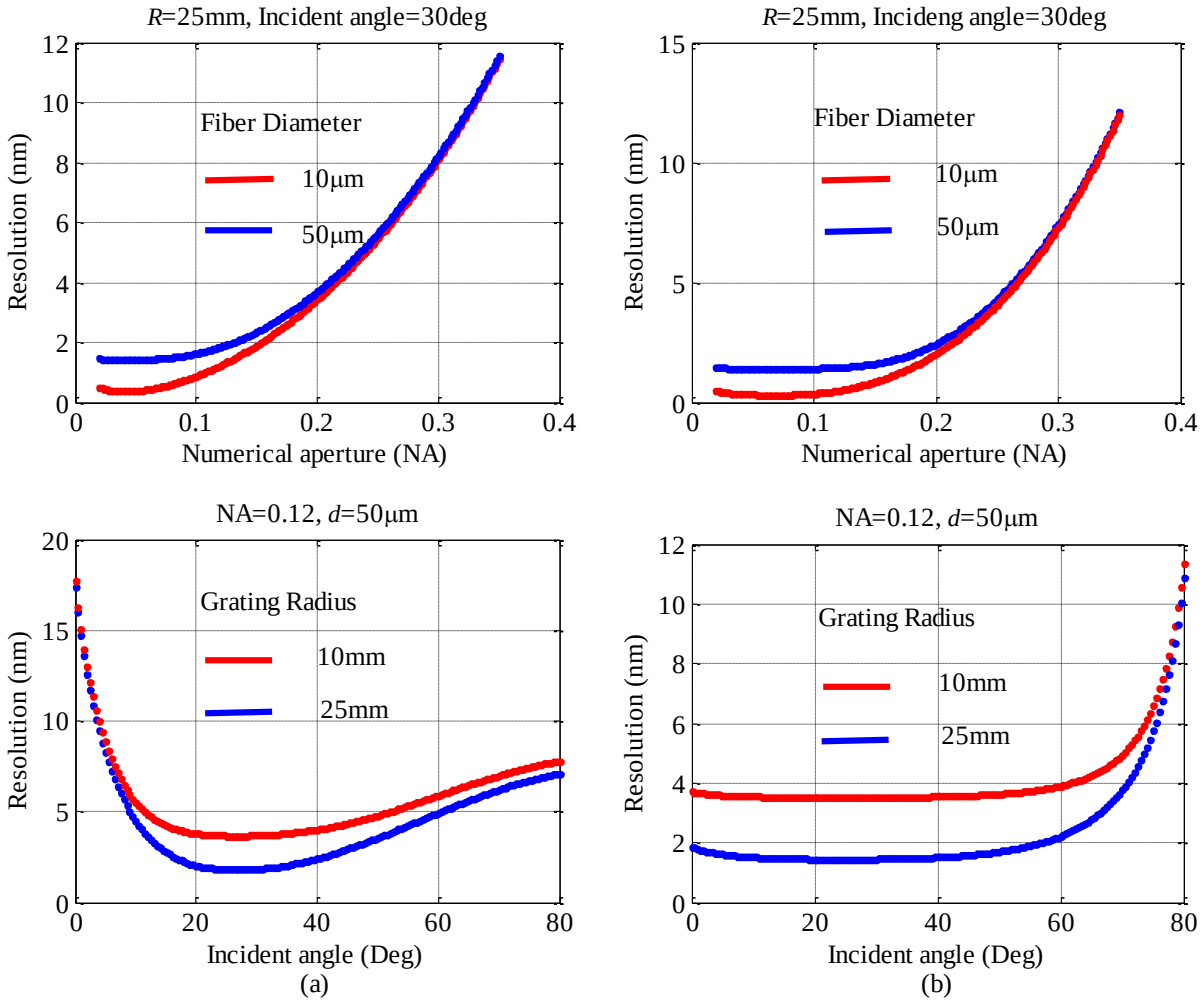


Figure 3.13: Total spectral resolutions when (a) the height *h* is close to the width *w*; and (b) the height *h* is much smaller than the width *w*.

Based on the results in Figure 3.13, the total spectral resolution decreases (improves) with the decreasing NA first (stage 1), then it reaches the minimum value at about 0.05 of NA. The exact position of the minimum value also depends on other design parameters. After the minimum position, the spectral resolution changes in the opposite way with further decreasing of NA (stage 2). Among the three contributions, $\Delta\lambda_{Entrance}$ is independent of NA, $\Delta\lambda_{Aberration}$ is proportional to NA and $\Delta\lambda_{Diffraction}$ is inversely proportional to NA. For stage 1, $\Delta\lambda_{Aberration}$ changes faster than $\Delta\lambda_{Diffraction}$, therefore the total spectral resolution is proportional to NA. After the minimum point, $\Delta\lambda_{Diffraction}$ changes faster than $\Delta\lambda_{Aberration}$, and the overall trend is inversely proportional to NA in

stage 2. Moreover, a large grating radius has a better spectral resolution due to the relatively smaller aberration. Therefore, to achieve good resolution, most concave gratings utilize large grating radii. Based on the requirements and limitations of specific applications such as handset systems for field applications, the above algorithms can be used to determine the optimum combination of different design parameters.

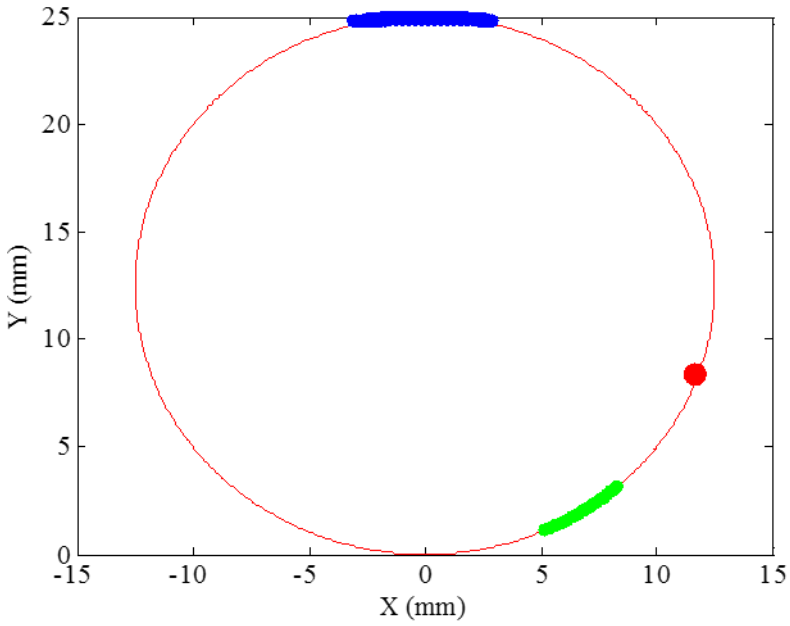


Figure 3.14: Horizontal focal curve of constant space concave grating (Incident 35deg, grating constant 700nm, radius 25mm, wavelength 550-650nm): Rowland circle (Red), grating (blue), focal curve (green)

3.2.3 Flat-field Concave Grating Design

The advantage of a Rowland concave grating has been introduced above. When the entrance slit is positioned on the Rowland circle, then light diffracted from the grating will also be focused on the Rowland circle. This configuration makes it easy to setup the system, but also leads to an important drawback of a Rowland concave grating—circular horizontal focal curve, which can be derived from the 2nd Taylor series coefficient F_{20}. If F_{20} is set to zero, then r_b in Eq. (3.14) gives the horizontal focal curve. Figure 3.14 shows the calculation result of the horizontal focal curve for wavelengths from 550nm to 650nm, from which we see that the focal curve (green) overlaps with the Rowland circle (red).

Since an image sensor typically has a planar photosensitive area, then using such a planar surface to collect diffraction light from a concave grating will broaden the spectrum and degrade the spectral resolution. Considering that a concave surface imager sensor is not common, then a flat-field concave grating (linear focal curve) is a better solution to this problem. Since all the

formulas given in Eq. (3.14) are based on a constant groove density, an efficient way to change the focal curve is to use the varied line space grating, by which, $\partial^i(mN\lambda)/\partial x^i \neq 0$. The power series coefficients can then be rewritten as [119]

$$F_{ij} = M_{ij} + mG_0\lambda H_{ij} \,. \tag{3.23}$$

In Eq. (3.23), the first term M_{ij} on right side is determined by the mounting parameters of the grating, such as the positions for entrance slit, grating and image plane. Note that H_{ij} is relevant to the groove density distribution along the x direction and it can be determined based on the optimization of the focal curve and aberration. Also, G_0 is the groove density at the grating center. To determine the structural information, each of the H_{ij} terms should be solved. Therefore, specific algorithms are derived for calculating the parameters of the flat-field concave grating.

A reference point Q_0 is picked from the horizontal focal curve first, and the wavelength corresponding to Q_0 is λ_0. To make the focal curve linear, the slope k_{QQ0} between the reference point $Q_0(\lambda_0\ x_0\ y_0)$ and any other point $Q(\lambda\ x\ y)$ should be constant. That is,

$$k_{QQ_0} = \frac{y - y_0}{x - x_0} = C \,, \tag{3.24}$$

where: $x_0 = r_{b0}\sin\beta_0,\ y_0 = R - r_{b0}\cos\beta_0;\ x = r_b\sin\beta,\ y = R - r_b\cos\beta$.

Applying the grating equation to Eq. (3.24), the slope between Q and Q_0 can be written as,

$$k_{QQ_0}(t) = \frac{r_{b0}\cos\beta_0 - r_b\cos\beta}{r_b\sin\beta - r_{b0}\sin\beta_0} = \frac{r_{b0}\sqrt{1-(mG_0\lambda_0 - \sin\alpha)^2} - r_b\sqrt{1-t^2}}{r_b\cdot t - r_{b0}(mG_0\lambda_0 - \sin\alpha)} \,, \tag{3.25}$$

where $r_b = \dfrac{\cos^2\beta}{\dfrac{\cos\beta}{R} - mG_0\lambda\cdot H_{20}} = \dfrac{R\left(1-t^2\right)}{\left(1-t^2\right)^{\frac{1}{2}} - R\cdot t\cdot H_{20} - R\sin\alpha\cdot H_{20}}$, $t = mG_0\lambda - \sin\alpha$.

The power series expansion of Eq. (3.25) in terms of t, gives

$$k_{QQ_0}(t) = g(t) = g_0 + g'(0)t + \frac{1}{2}g''(0)t^2 + \frac{1}{6}g'''(0)t^3 + ... \tag{3.26}$$

To guarantee a constant slope, which means keeping only the g_0 term, then the sum of all other terms is set to zero. The idea of this design is to find a specific H_{20} which makes the sum of all "t" terms in Eq. (3.26) close to zero. To avoid the effect of "pole point" and to make the focal curve stable, the reference wavelength should be chosen out of the wavelength band.

Using the Taylor's expansion formula, the best H_{20} is found to be as in Eq. (3.27). Details of this calculation is given in Appendix I.

$$H_{20} = \frac{\sqrt{1-t_0^2}}{RmG_0\lambda_0} = \frac{\sqrt{1-\left(mG_0\lambda_0 - \sin\alpha\right)^2}}{RmG_0\lambda_0}. \tag{3.27}$$

H_{20} is dependent on the groove density G_0, grating radius R, reference wavelength λ_0 and incident angle α. Figure 3.15 shows the simulation results of the variation of H_{20} with incident angles from 30° to 80° at two reference wavelengths. As shown in the simulation results, H_{20} decreases with increasing reference wavelength and increases with incident angle. When a different incident angle is used, to achieve linear focusing, the groove density distribution must be modified accordingly.

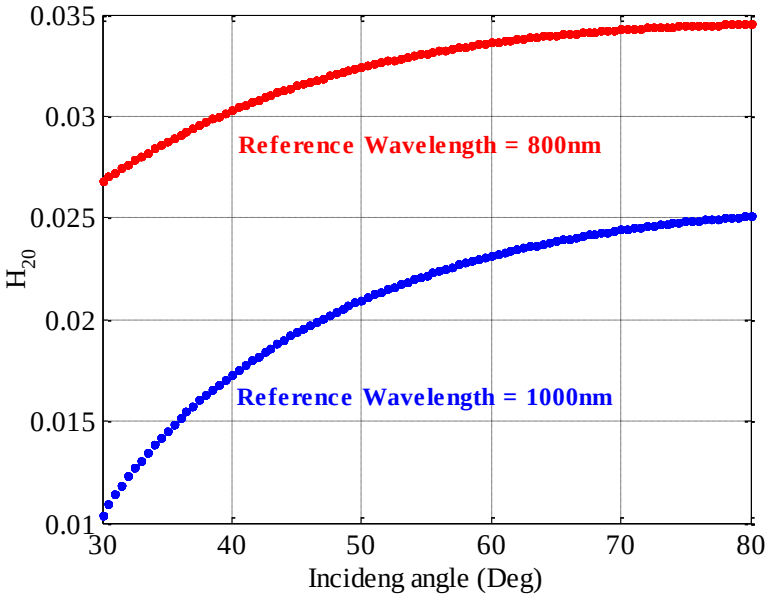


Figure 3.15: Dependence of H_{20} on the incident angle α and reference wavelength λ_0

With H_{20} given in Eq. (3.27), the modified horizontal focal curve can be calculated by Eq. (3.23). To verify the importance of the reference wavelength, Figure 3.16 shows the calculation results of the sum of several high order terms in Eq. (3.26) and the modified horizontal focal curve with different reference wavelengths. As shown in the simulation results, when the reference wavelength lies within the target wavelength band, the "pole effect" in Eq. (3.25) causes the divergence of the focal curve. In contrast, when the reference wavelength is selected out of the target wavelength band, a linear focal curve can be achieved [Figure 3.16(b)].

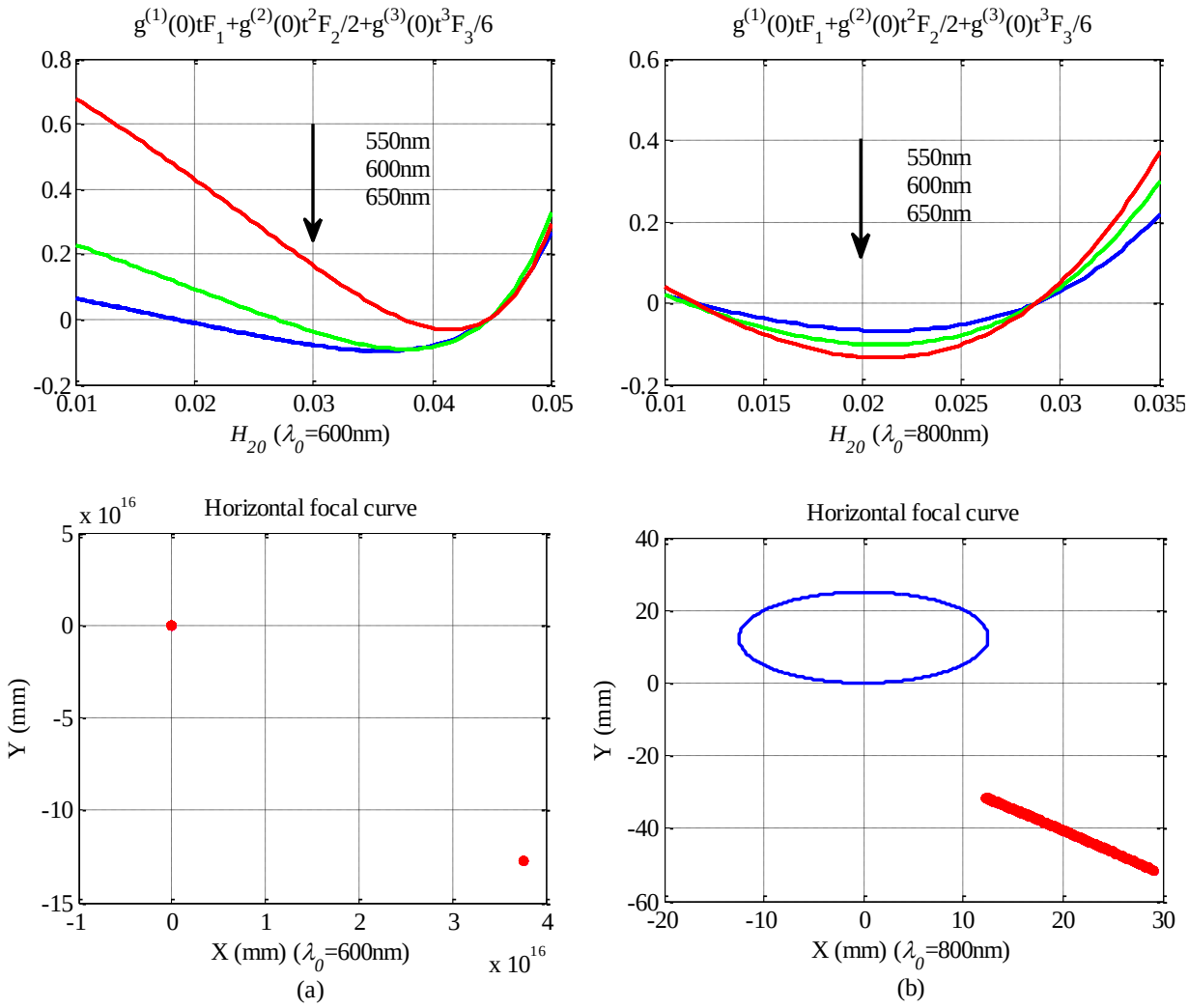


Figure 3.16: Sum of the first three high order terms in Eq. (3.26) and the modified linear focal curve (Red) corresponding to the original circular focal curve (blue) with for wavelength band from 550nm to 650nm (Grating constant 700nm, incident angle 35deg): (a) Reference wavelength=600nm; (b) Reference wavelength=800nm

With the linear horizontal focal curve, Figure 3.17 shows an example of a concave grating based system, designed for a wavelength band from 550nm to 650nm. A 4mm grating radius is used for a miniaturized system. To make the system more compact, a mirror was used to reflect the diffraction beams to detector1.

In addition, a baffle was added between the entrance slit and Detector1 to avoid direct illumination from the entrance to the detector. The entire system is marked by the dotted (red) rectangular area, with size of ~1mm x 4mm x 3.8mm. The system shown in Figure 3.17 is an example of a miniaturized system. To design a grating system, the grating radius can be selected according to the required specifications and available fabrication processes.

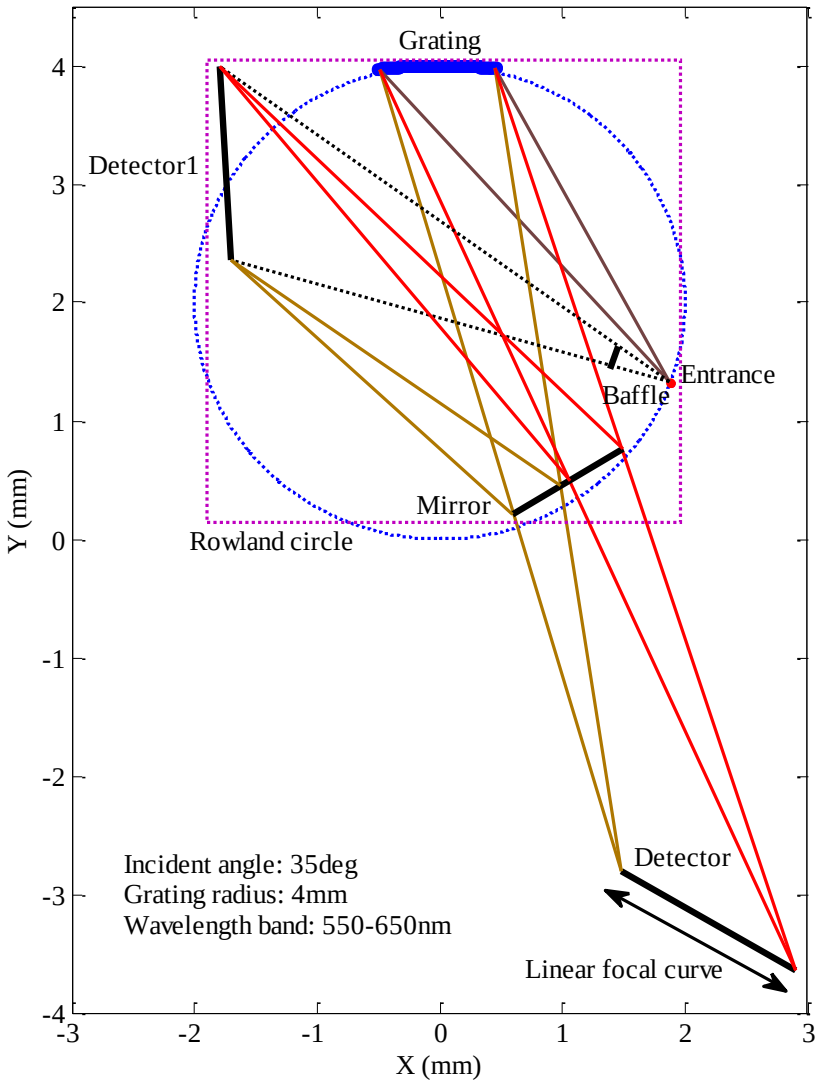


Figure 3.17: Optical diagram of concave micro-grating spectrometer system (Reference wavelength=900nm)

3.3 Concave Grating Fabrication and Characterizations

3.3.1 Grating Fabrication Techniques

Given the grating information (groove density and groove profile), a grating can be fabricated by different techniques. However, because the grating structures have to be formed on a concave substrate, fabrication of a concave grating is more challenging and expensive compared to a planar grating.

The fabrication of gratings usually starts from the fabrication of a master grating, which can then be replicated to produce many subsequent product gratings. The replication process significantly reduces the cost of gratings. To date, two commonly used techniques for master gratings fabrication are the mechanical ruling and holographic method.

The mechanical ruled gratings are the first generation diffraction gratings, in which the grooves are mechanically ruled individually with a diamond tip either on planar or concave substrates, coated with a thin film of metal. The shape of the diamond tip determines the groove profiles, for example, laminar grating or blazed grating. During fabrication, the diamond is controlled by a ruling engine, which is the most vital component in the ruled grating manufacturing, and it should be free from vibration, temperature and atmospheric variations. Since each groove is ruled individually, the mechanical ruling process is time consuming.

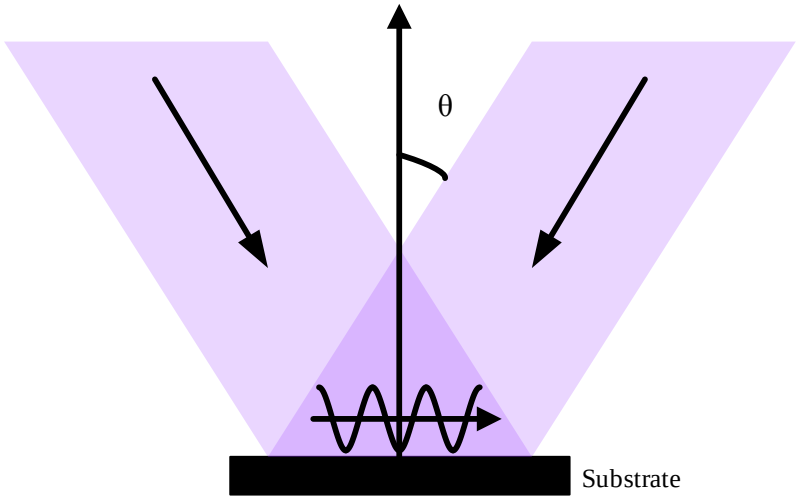


Figure 3.18: Schematic of holographic grating fabrication

Distinct from the mechanical ruling method, the holographic method relies on the standing wave pattern formed by two sets of coherent monochromatic light sources with equal intensity and wavelength. A planar or concave substrate coated with a thin layer of photoresist is exposed to the standing wave pattern, the intensity of which varies from zero at the destructive interference point to the maximum at the constructive interference point. The grating profile can be formed by developing the exposed substrate in a relevant developer of the photoresist. According to the type of photoresist, either the exposed or unexposed regions can be removed in the developer, thus forming a sinusoidal grating profile. The width of the groove can be calculated by Eq. (3.28) [119], where λ is the wavelength of the light source, and θ is the half angle between the two beams (shown in Figure 3.18),

$$d = \frac{\lambda}{2\sin\theta}. \tag{3.28}$$

A photoresist is sensitive to light with a specific wavelength. Once the photoresist is chosen, then the wavelength of the light source is also fixed. According to Eq. (3.28), the grating constant

can be modified by adjusting the angle between the two interference beams. Since all grooves are formed simultaneously, the time it takes to fabricate a holographic grating is much shorter than that for a ruled grating. However, owing to the intensity distribution of the interference pattern, the holographic gratings usually have sinusoidal grooves, and extra treatment is required to make the holographic grating blazed, for example, by ion etching. In contrast, it is relatively easier to obtain blazed gratings by mechanical ruling.

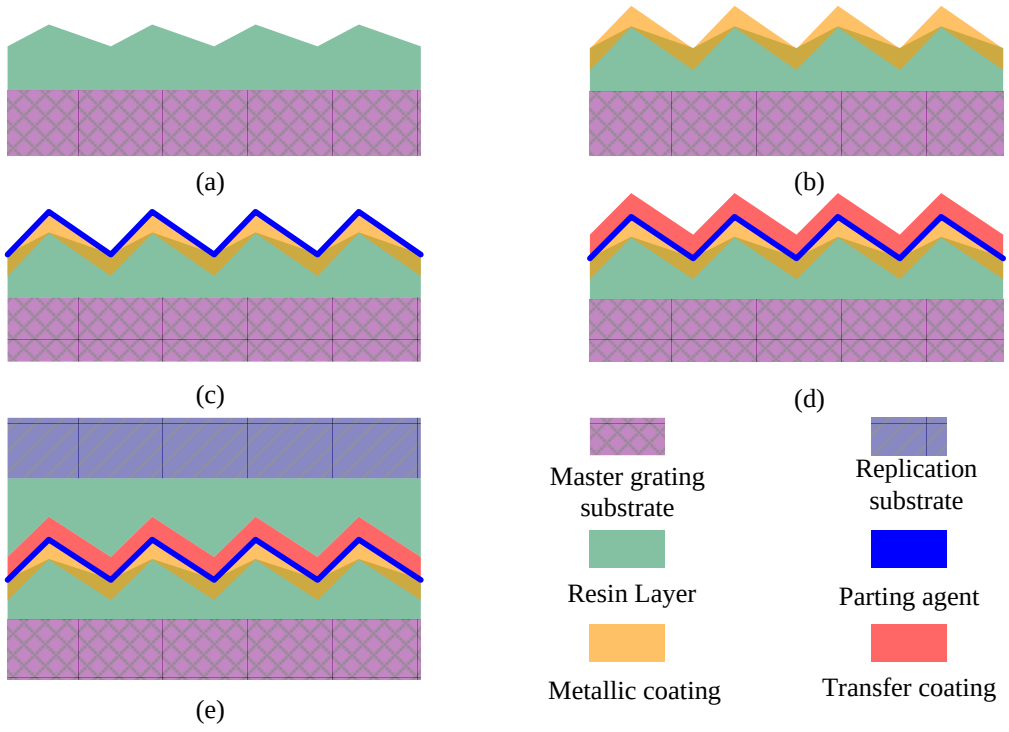


Figure 3.19: Replication of a master grating. The grating structure (e) is from [119].

After fabrication, the master grating is replicated through the processes shown in Figure 3.19. The replication process starts from a thin reflective metallic layer coating [Figure 3.19(b)], which is optional if the mast grating will not be used as a product grating. After that, a very thin layer of parting agent is applied [Figure 3.19(c)]. This parting agent functions as a separation between the master grating and the replication grating. Then a second layer of metal, working as the reflective layer of the product grating is coated. Since this layer will be transferred to the product grating, it is called the transfer coating [Figure 3.19(d)]. The last two layers are a resin layer and a replication substrate [Figure 3.19(e)]. Initially, the resin layer is a liquid adhesive. Since its purpose is to hold the grooves and maintain their profiles, the resin layer has to be cured to harden, and also gluing to the replication substrate. Finally, the product grating is separated from the master grating through

the parting agent layer, with the shape of the groove profiles of the master grating reversely duplicated on the product grating.

3.3.2 Concave Grating Fabrication

The two fabrication and replication processes discussed above are commonly used in industry for grating products. Concave gratings fabricated by the mechanical ruling method are very expensive, especially for custom gratings with variable line space. Owing to advantages in system miniaturization, there have been numerous efforts in academia towards the fabrication of concave gratings, using different micro-fabrication processes, such as the deep x-ray lithography [52] and UV nano-imprint lithography [51]. Unfortunately, these fabrication processes are also expensive.

This research is to build a low-cost and compact system for environmental application. To reduce the cost, the constant line space concave grating is considered, rather than the costly varied line space concave grating. Consequently, the horizontal focal curve of the system is circular, as discussed in section 3.2.3. As for the fabrication, the low-cost holographic method with a single illumination source is employed.

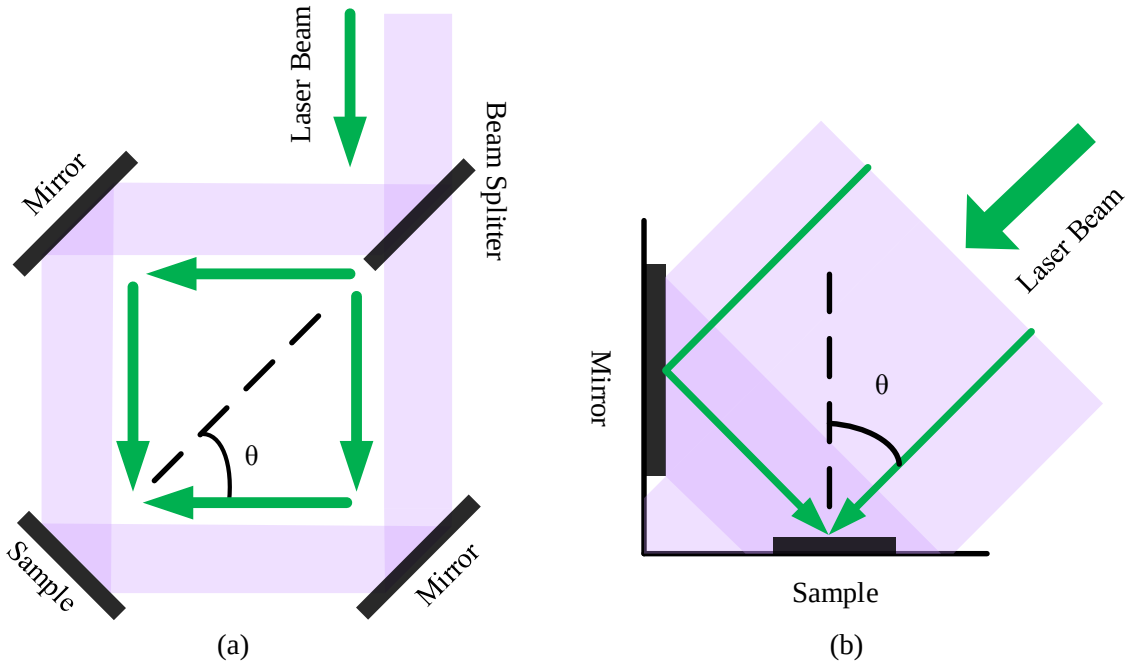


Figure 3.20: Optical setup of the holographic fabrication: (a) Common setup; (b) Setup in this design

Figure 3.20 shows two types of optical setups for the holographic fabrication of a grating. Setup (a) is a common way to implement the holographic method. A monochromatic light (usually from a laser) is divided into two beams by a beam splitter, with 50% reflected and 50% transmitted. The

two beams are then reflected by two mirrors mounted on both sides of the beam splitter. The reflected beams finally meet and generate the interference pattern on the plane where the grating substrate is mounted. The grating constant in this setup can be calculated by Eq. (3.28), and its value can be adjusted by rotating the two mirrors to change the intersection angle between the two beams. To obtain high quality interference patterns, the two reflected beams should have equal intensity and the two mirrors should be symmetrically positioned on both sides of the beam splitter. An advantage of this setup is that a relatively larger interference area can be generated. Unfortunately, owing to the complex setup, poor interference patterns were obtained. Grating structures were failing to form on the grating substrate, even when the substrate was a flat silicon wafer.

Alternatively, the much simpler setup (b) was used, which was arranged on an optical L stage. A mirror was mounted on the vertical arm of the L stage and the grating substrate was placed on the horizontal arm of the L stage. During fabrication, a collimated laser beam of certain diameter (~10 cm) was incident on the L stage, covering both the grating substrate and the mirror. The beam reflected from the mirror interfered with the un-reflected beam, generating the interference pattern of desired pitch that depends on the tilt angle of the laser beam to the L stage.

Compared with setup (a), the advantage of the second setup is that better interference patterns were obtained. However, an important limitation of this setup is that the available area (overlapping area) of the interference pattern is related to the tilt angle of the L stage, which then determines the grating constant. A UV laser source (325nm) was used in this system. For this specific wavelength, the available overlapping area was ~3-4cm for grating constants ~300nm, reduced to ~2cm for grating constant ~600nm. An even narrower overlapping area was achieved <1cm when the grating constant was ~900nm. Thus, there is a trade-off between the grating constant and the available grating area. For a larger grating area, the grating constant has to be reduced. However, as simulated in section 3.2.1, grating constants between 550nm and 1000nm are preferred for the proposed system.

Plano-concave ($18) and plano-convex ($20) lenses from Thorlabs were used as the substrates. Figure 3.21(a-b) shows the images of samples before fabrication. An important advantage of the lens is that one side is planar, so the lens can be firmly mounted on a regular chuck of the spinner, and high spin speed can be applied for a more uniform photoresist layer. In addition, the lens is

designed for optical use, so the surface shape and smoothness are better than of a watch glass (See fabrication challenges in Appendix II).

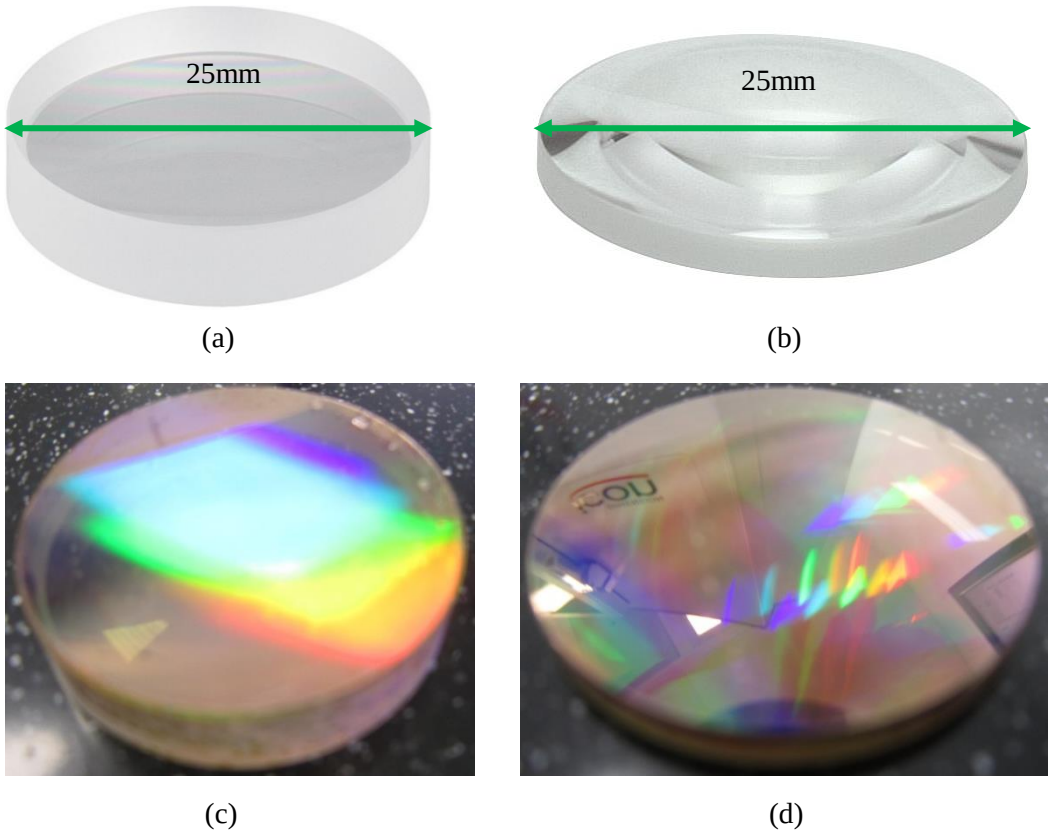


Figure 3.21: Glass substrate from Thorlabs: (a) Plano-concave lens (LC1258); (b) Plano-convex lens (LA1257); (c) Grating on Plano-concave lens; (d)Grating on Plano-convex lens

Grating structures were fabricated on the lenses as follows: the lens was spin-coated with a thinned photoresist (S1808) with spin rate of 5000rpm for 30s. Considering the different thermal conductivities of silicon and glass, the photoresist was soft baked on a hot plate at 90°C for ~4mins, and no photoresist accumulation was observed. The baked lens was then exposed to the interference pattern for a period of 40s. Then, the exposed sample was developed in the CD30 solution, and development time of the silicon-based planar grating was used as a reference. After development, the photoresist was finally hard baked at temperature of 120°C for ~2mins. To make the surface reflective, the sample was sputter-coated with 2.5nm Cr and 20nm Au. Figure 3.21 (c) and (d) show diffraction from the two lens-based diffraction grating.

To test the efficiency of this process in producing periodic grating structures, the surface of the samples was inspected with a Scanning Electron Microscope (SEM). The SEM image of the

concave glass surface is shown in Figure 3.22, from which we see the grating structure, with a grating constant of ~980nm.

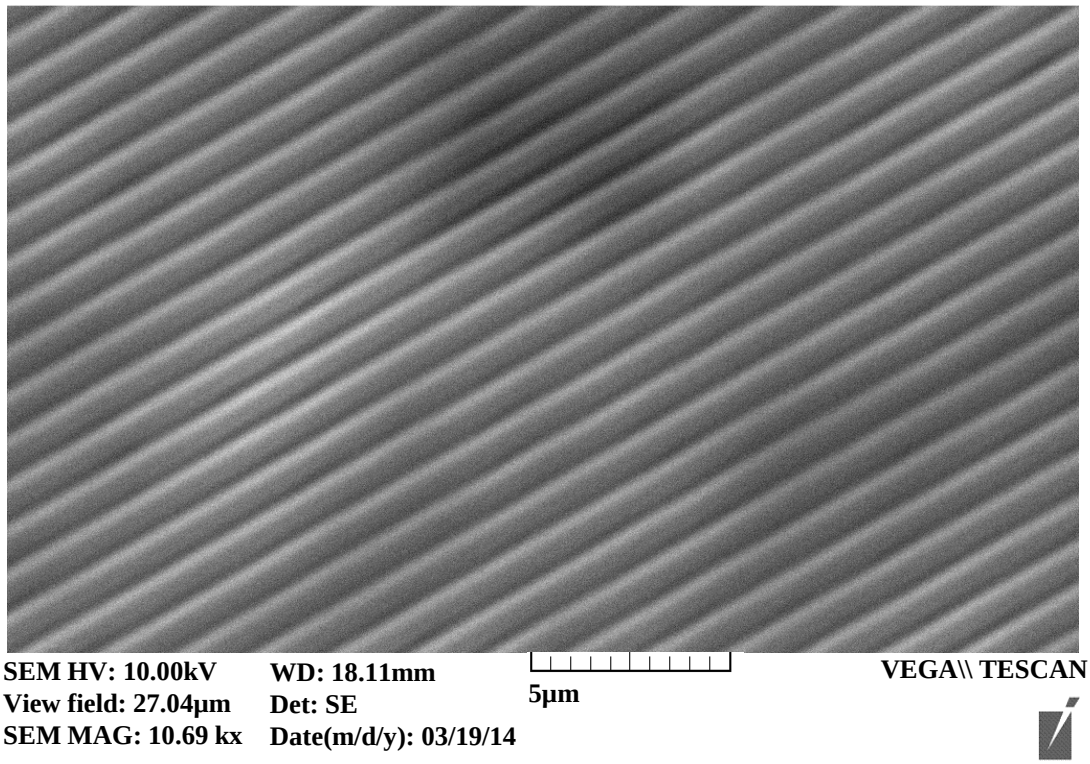


Figure 3.22: SEM image of the concave grating surface

3.3.3 Concave Grating Measurements and Comparison to Theory

According to the SEM measurements, grating structures have been successfully produced on the plano-concave lens. The next step is to verify the optical characteristics. The primary purpose of this grating is the dispersion of light. For a concave grating, the focusing property is also very important. To verify the dispersion and focusing characteristics, an optical test system was built. Instruments and components required for this setup are a light source, light transmission optics, and a detector. However, restricted by the Rowland configuration and grating radius, it is difficult to carry out this experiment.

Figure 3.23 shows the schematic diagram of a concave grating system with the Rowland configuration. As discussed before, the incident source, concave grating, and detector are all on the Rowland circle. The incident light was transmitted by a fiber, so the entrance slit was one end of the fiber. A commercial camera was used to capture image from the grating. As shown in Figure 3.23, diameter of the Rowland circle was only 38.6mm, the narrow space between the grating and

the focusing point of the light makes it impossible to assemble a large packaged camera into the small Rowland circle.

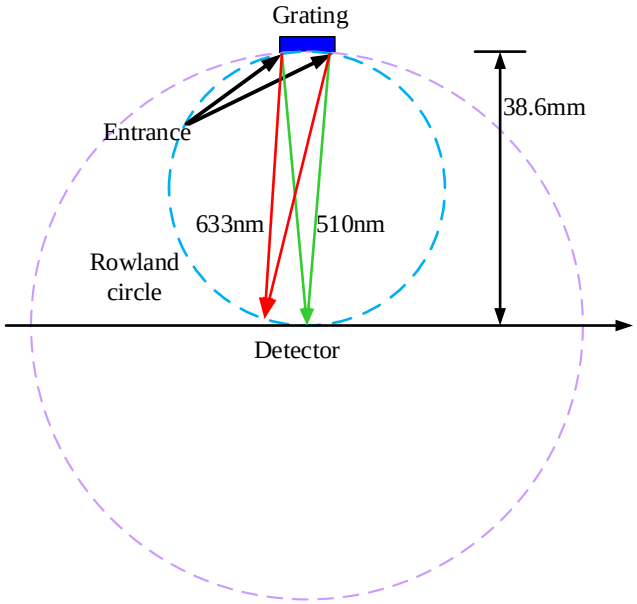


Figure 3.23: Optical diagram of the Rowland configuration with source, grating and detector

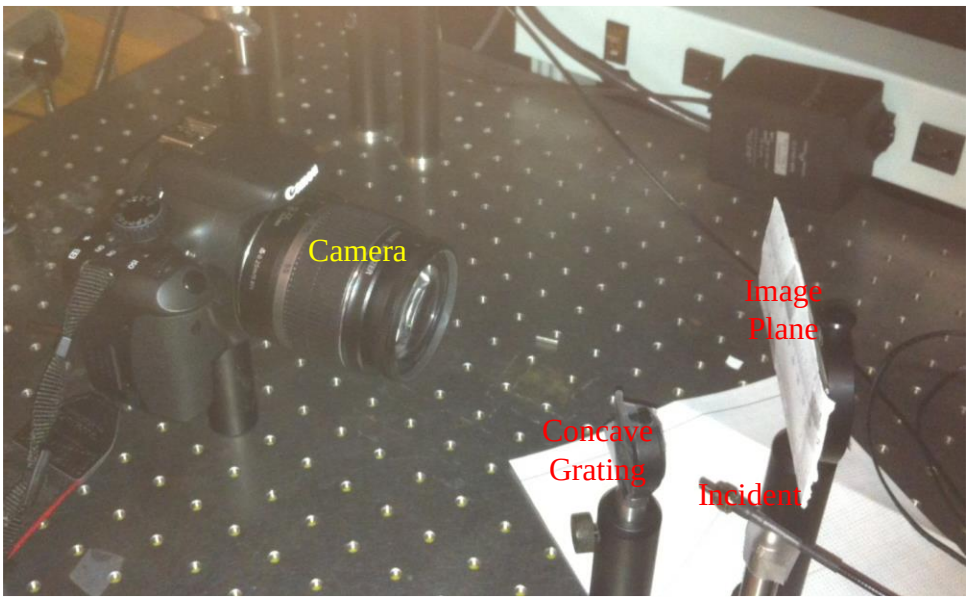


Figure 3.24: Experimental setup of the concave grating test

Therefore, the test system was finally set as shown in Figure 3.24. A white image plane was placed at the focusing point of the Rowland circle. The commercial camera was used to take the image focused on the image plane. First, the dispersion property was measured with a whit light as the incident source. Figure 3.25 shows the image taken on the image plane by the camera, from

which the 0th order reflection (white) and the 1st order (rainbow) diffraction are well displayed, proving the dispersion property of this custom concave grating.

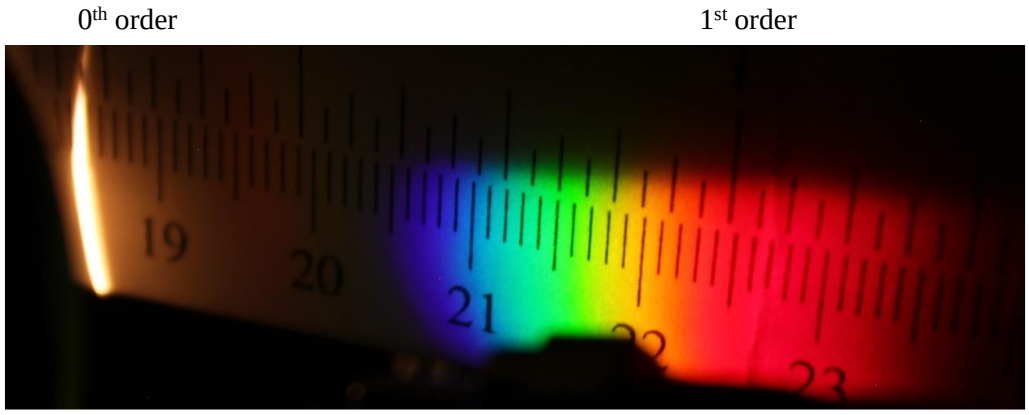


Figure 3.25: Dispersion of the concave grating of a white light

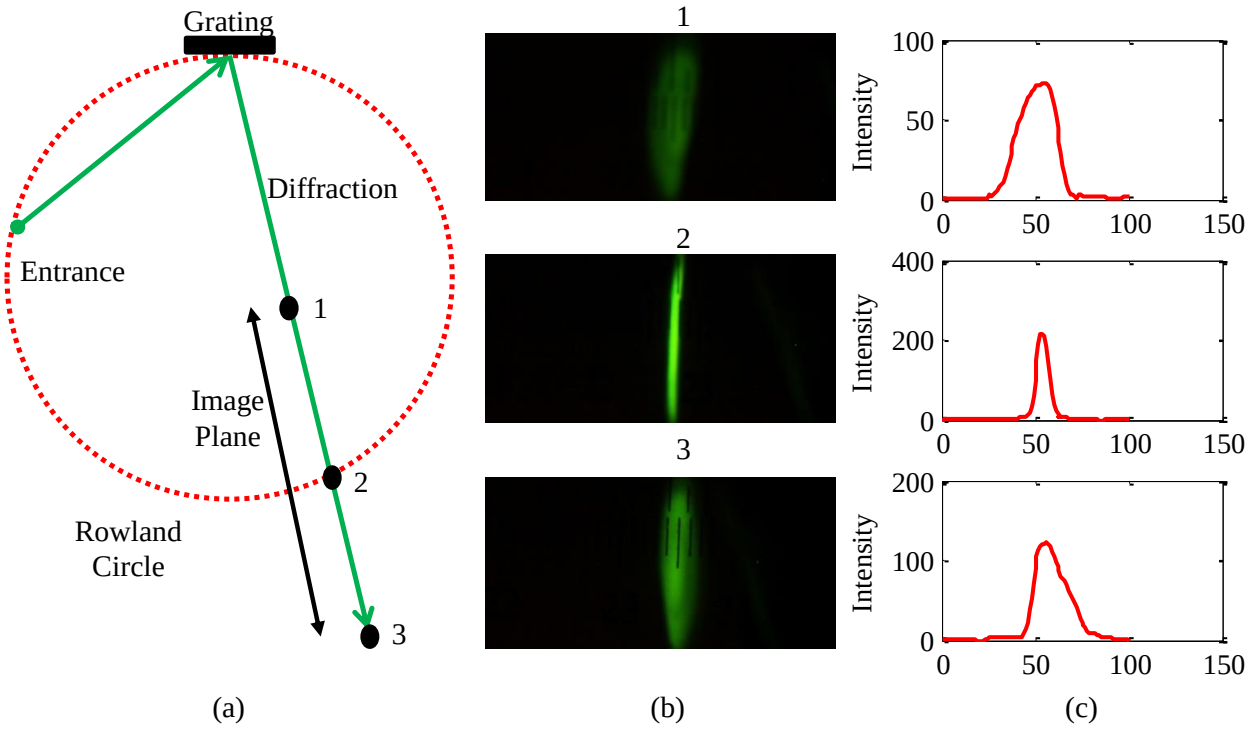
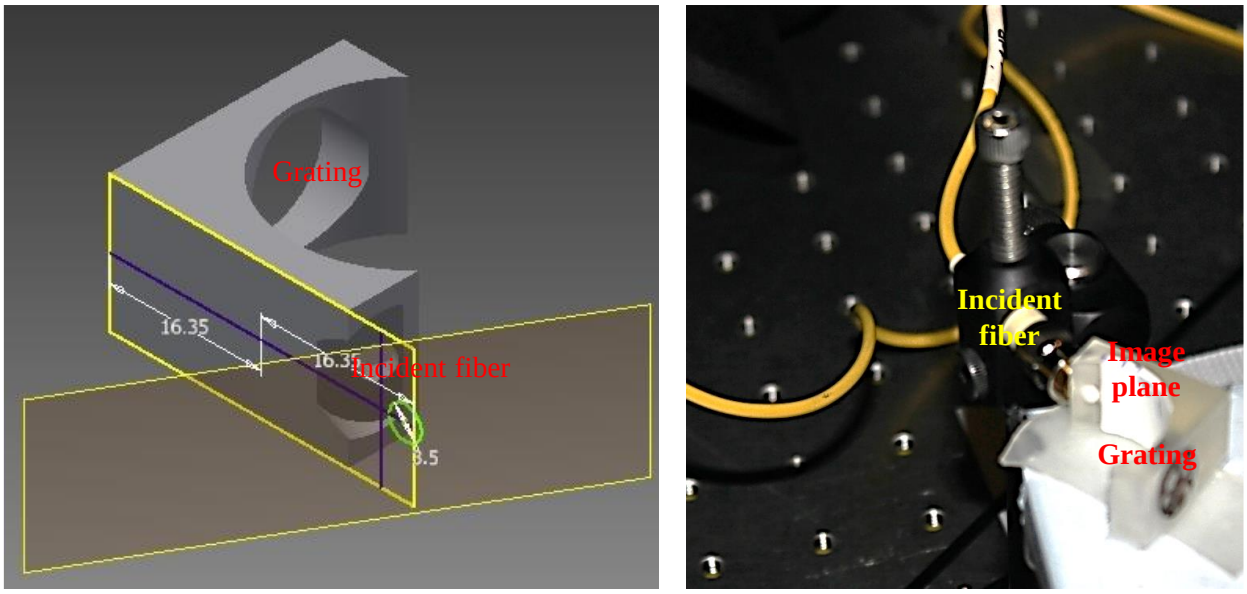


Figure 3.26: (a) Schematic of the focusing property measurement setup; (b) Images taken off and at the Rowland circle; (c) Intensity distribution of images

Second, the focusing property was measured. To verify the focusing property, the incident light has to be narrow-band. Therefore, a filter with 10nm bandwidth and center wavelength of 560nm was used. Since the diffraction beam is only focused on the Rowland circle, the image plane was moved along the diffraction direction, and images were taken at the Rowland circle and two arbitrary points off the Rowland circle. Figure 3.26(a) shows the schematic setup of the

measurement. The measured results and the intensity distributions are given in Figure 3.26(b-c). According to the measured results, the image taken at the Rowland circle (#2) is narrower and brighter than images taken from positions out of the Rowland circle (#1 and #3). This result proves that the diffraction light is more focused on the Rowland circle than at other positions.

The dispersion and focusing properties tested above have verified that the custom concave grating can perform the basic concave grating functions. Other performances of the concave grating, such as its spectral resolution, are also important. Since the horizontal focal curve of the concave grating overlaps with the Rowland circle, the setup shown in Figure 3.24 cannot be used for this measurement, because it is difficult to manually move the image plane to the exact Rowland circle. To measure the spectral resolution, a 3D model was designed to mount more precisely the grating, entrance fiber and image plane on the Rowland circle.



(a) (b)

Figure 3.27: (a) 3D model designed by Autodesk inventor; (b) Optical setup with the 3D model

The 3D model was designed with Autodesk inventor [Figure 3.27(a)] and fabricated using a 3D printing machine. Figure 3.27(b) shows the grating, incident fiber and image plane along the Rowland circle. Again, a commercial camera was used to take the image on the image plane. Table 3.1 lists the specifications of all components used.

The measurement of the spectrum of a 633nm laser is shown in Figure 3.28(a). The measured data were fitted with a Gaussian fitting[1]. Based on the FWHM of the measured spectrum and the

[1] Curve fitting tool box in MATLAB

bandwidth of the incident laser, the spectral resolution of the system shown in Figure 3.27(b) was calculated to be ~1.6nm. With the specifications given in Table 3.1, Figure 3.28(b) shows the result from simulation of the variation of spectral resolution with incident angles using Eq. (3.22). The simulation result of 1.4nm was found to be in good agreement with the measurement result of 1.6nm for FWHM.

Table 3.1: System specifications of the spectral resolution measurement

Component	Specifications	
Grating substrate (Thorlabs LC1439)	Size	½″
	Radius	25.7mm
	Grating constant	~540nm
Incident light	Wavelength	633nm
	Bandwidth	1nm
Entrance slit	Incident angle	30deg
	Entrance diameter	8μm
	Numerical aperture	0.14

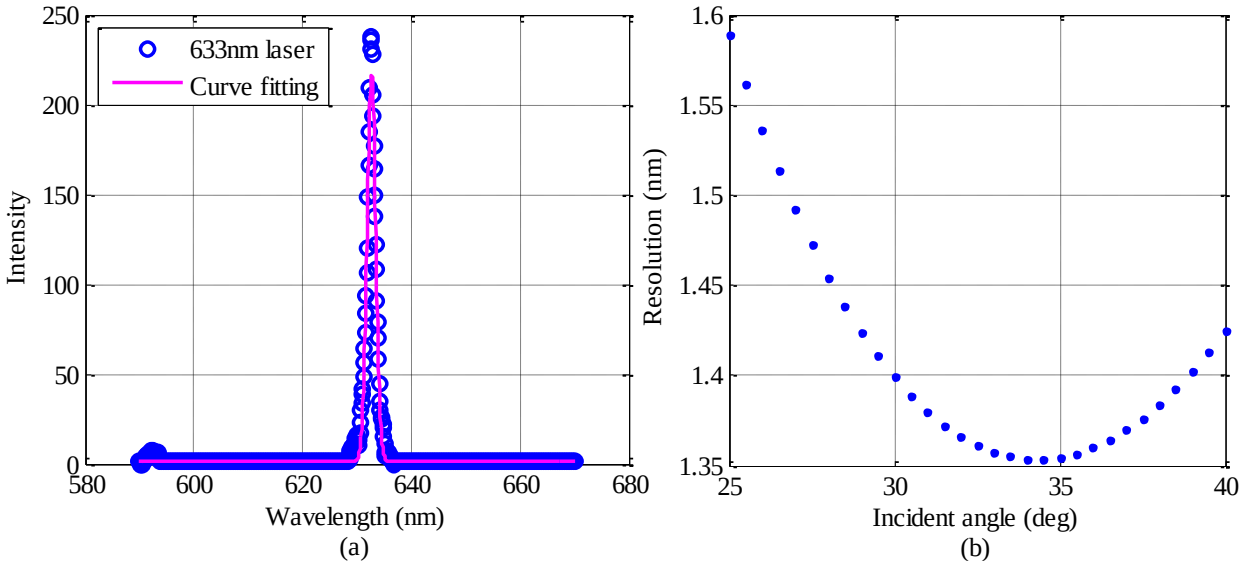


Figure 3.28: Spectrum resolved by the concave grating

3.3.4 Future Improvements

Based on the design and measurements, it was concluded that the grating fabricated by the custom low-cost holographic method can perform the basic functions of concave gratings, and a fine spectral resolution was achieved. One problem for application of this custom concave grating is

the circular horizontal focal curve. In principle, there are two possible ways to improve the performance of the concave grating. The first one is to modify the focal curve by varying the line space of the grating, for example, the flat-field concave grating. However, fabrication of a varied line space grating is more sophisticated and costly than that of a constant line space grating.

On the other hand, the utilization of the concave grating can also be improved if the system is operated as a monochromator. Typically in a monochromator, the output wavelength is adjusted by rotating the planar grating. This rotation changes the incident angle of the collimated light on the planar grating, which then changes the output wavelength at the exit slit accordingly. For a concave grating, limited by the Rowland configuration, the incident angle can only be adjusted along the Rowland circle. As a result, a motor can be designed to automatically move the detector along the Rowland circle.

3.4 Summary

In this chapter, diffraction theory with the purpose of creating diffraction gratings for wavelength selectors was reviewed. Several properties of a diffraction grating were analyzed. To minimize the overall size and complexity of the wavelength selector, a concave grating was designed. The design of the concave grating was accomplished through simulation of diffraction efficiency and spectral resolution, from which the optimum combination of the design parameters can be determined for a target specification. In addition, considering the circular horizontal focal curve, an algorithm was proposed for the design of a flat-field concave grating. The concave grating was finally fabricated on plano-concave/convex lenses by a low-cost custom holographic method. The functions of the grating were tested and for the evaluation of the spectral resolution, a 3D model of the test setup was designed and fabricated. A 1.6nm spectral resolution was achieved, which is very close to the calculated result of 1.4nm.

Chapter 4

Time-Gated Single Photon Avalanche Photodiode (TG-SPAD)

In this chapter, the theory, design and characterization of the detection system is described. Since the proposed system is for Raman spectroscopy application, the detector for this system must be very sensitive. As discussed in chapter 2, to design a low-cost and compact detector for low light intensity detection, the silicon-based single photon avalanche diodes (SPADs) was selected.

4.1 SPADs-Review and Theory

SPADs have been extensively explored for the detection of low intensity light. The implementation of SPADs in complementary metal-oxide-semiconductor (CMOS) technologies has boosted the application of the CMOS SPADs, owing to its integration capability with CMOS control electronics [120]-[122]. Because of single photon sensitivity and fast timing response, SPADs have been introduced to numerous fields of time-correlated single photon counting (TCSPC) and time-gated detection applications, including fluorescence lifetime imaging (FLIM) [123], near infrared spectroscopy (NIRS) [31], [124], and light detection and ranging (LIDAR) [125]. A SPAD can be operated in either free running or time-gated (TG) modes. The fast gating capability makes the CMOS SPAD a low-cost alternative to ICCDs and PMTs. Recently, CMOS TG-SPADs have been proposed for fluorescence rejection in Raman spectroscopy [126]-[129].

A SPAD is essentially a *pn* junction biased above the avalanche breakdown voltage (V_{BR}), the so-called Geiger mode. Controlled by an external quenching circuit, the single photon detection process of the SPAD can be divided into four phases. These are carrier generation by a photon, internal carrier multiplication, quenching of the voltage across the SPAD, and recharge. Figure 4.1

shows the four phases involved in a photon detection process reflected superimposed on the IV characteristic of the SPAD.

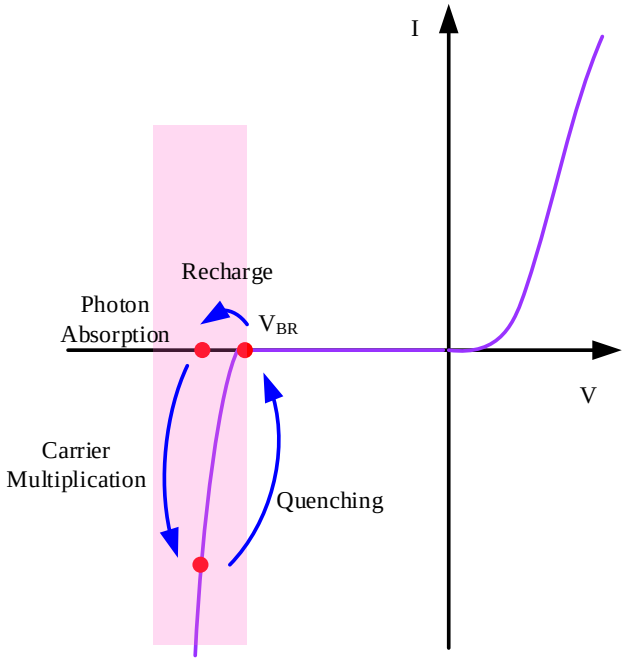


Figure 4.1: Current-voltage (IV) characteristic of avalanche photodiode operated in Geiger mode

Before photon detection, the SPAD is biased above the breakdown voltage (V_{BR}), and this overdrive voltage is named as the excess bias (V_{ex}). The detection cycle of the SPAD begins when an absorbed photon results in the generation of an electron-hole pair of free carriers. In the second phase, the photon-generated carriers are accelerated by the high electric field and multiplied in the depletion region. As discussed in [130], basically two processes are involved in this phase. The first process refers to impact ionization induced by the high electric field, which triggers the carrier multiplication process that creates an avalanche of secondary carriers. The ionization probability strongly depends on the electric field, which can be adjusted by applying different V_{ex}. The second process is an internal quenching. The ionization process increases the local current density, which causes a larger voltage drop across the space charge resistor and in turn weakens the ionization process. When the two processes are balanced, a current pulse in the milliampere range is then delivered to the external control circuit which senses and quenches the avalanche.

The fast quenching of the avalanche process is important for the operation of the SPAD, because it protects the diode from overheating and related damages. After fully quenched, the SPAD is recharged back to the Geiger mode. The entire detection cycle is shown in Figure 4.1. The external control circuit plays a dominant role in the detection cycle. The SPAD control circuit

is designed to match a specific application. For instance, the SPAD can be designed in a pixel array or as a discrete component, operated in free running or time-gated mode. Each SPAD application has different requirements, and the quenching and recharge circuits of the SPAD must be designed accordingly. The parameters commonly used to characterize the performance of a SPAD are now described.

(1) *Dead time*: As depicted in Figure 4.1, during the time when a SPAD undergoes photon absorption, carrier multiplication, quenching, and recharge, the SPAD is not able to detect a subsequent photon. Dead time is defined as the time it takes to complete a detection cycle, which determines the maximum counting rate of a SPAD. Therefore, fast quenching and recharge circuits are required for high-speed applications.

(2) *Dark count rate* (DCR): DCR is defined as the number of counts per second when the SPAD is in dark. A major source of DCR in a CMOS SPAD at room temperature is the thermal generation of free carriers, and the generation rate is also related to the ionization probability. Therefore, DCR depends on temperature, excess bias, and size of the active area of the SPAD. DCR determines the minimum incident photon rate that can be detected. Detailed characterization of the DCR of the designed SPAD pixel is given in section 4.3.2.1.

(3) *Photon detection efficiency* (PDE): When a light source with certain intensity is incident on the detector, only a portion of the incident photons can be detected due to several reasons, including surface reflection, photon absorption before reaching the depletion region, absorption coefficient of the material, and triggering probability of an avalanche. PDE is defined as the ratio of the number of voltage pulses detected to the number of total incident photons. PDE is the most important parameter to evaluate the detection efficiency of a SPAD, and it is a function of the incident wavelength and excess bias. PDE of a SPAD corresponds to the external quantum efficiency of other types of photodetectors.

(4) *Afterpulsing probability* (AP): During the quenching phase, a large number of carriers flow through the depletion region. Some carriers are trapped in deep energy levels within the depletion region and released later to trigger a second detection cycle that is not initiated by photon absorption. The pulses generated by the released carriers from traps are named as the afterpulsing in SPAD. Afterpulsing probability is correlated to the avalanche current and its duration, thus, to the amount of charge of the avalanche pulse. Therefore, AP is

proportional to the parasitic capacitance of the photodiode and the quenching time. The latter can be optimized by specially designed external control circuits.

(5) *Fill factor* (FF): To achieve fast quenching and resetting, complex control circuits are designed for SPAD pixels. In active SPAD pixels, the area of the control circuits is often comparable to or larger than the optically active area of the SPAD. Fill factor is defined as the ratio of the SPAD active area to the total pixel area. For better usage of the chip area, a high FF is preferred.

DCR and PDE are strongly determined by the fabrication process, although some freedom is available for the designs of the control circuits in SPAD pixels. Optimized designs of quenching and recharge circuits mostly target reduction of the dead time and afterpulsing probability, and increase in fill factor. Two types of SPADs control circuits — free running and time-gated will be discussed next.

4.1.1 Free Running Operation

Passive quenching and recharge is the simplest approach for design of a SPAD in free running mode, which usually contains a quench resistor (R_Q) connected in series with the photodiode which is biased at (V_{BR}+V_{ex}). To efficiently quench the avalanche current in a very short time, R_Q is very large, usually in the range of kΩ to MΩ. Consequently, even a low avalanche current can result in a high enough voltage drop across R_Q. This voltage drop forces the bias of the SPAD to reduce below the breakdown voltage V_{BR} and then quench the avalanche process. The time it takes to fully quench the avalanche current is determined by R_Q, and the dynamic resistance (R_D) and capacitance (C_D) of the diode.

Figure 4.2(a) shows a simple SPAD circuit with passive quenching and recharge. The equivalent quenching and recharge circuits are given in Figure 4.2(b-c) respectively. The external biases are applied to the anode (V_A) and the quench resistor (V_C), and photon detection is sensed from the variations of the cathode potential [$V_O(t)$]. With this bias condition, the excess bias equals to

$$V_{ex} = V_C - V_A - V_{BR}\,. \tag{4.1}$$

During passive quenching [Figure 4.2(b)], $V_O(t)$ can be calculated with Kirchhoff's law for the current at the cathode node

$$\frac{V_C - V_O(t)}{R_Q} = \frac{V_O(t) - V_A - V_{BR}}{R_D} + C_D \frac{\mathrm{d}V_O(t)}{\mathrm{d}t},\ V_O(t=0) = V_C\,. \tag{4.2}$$

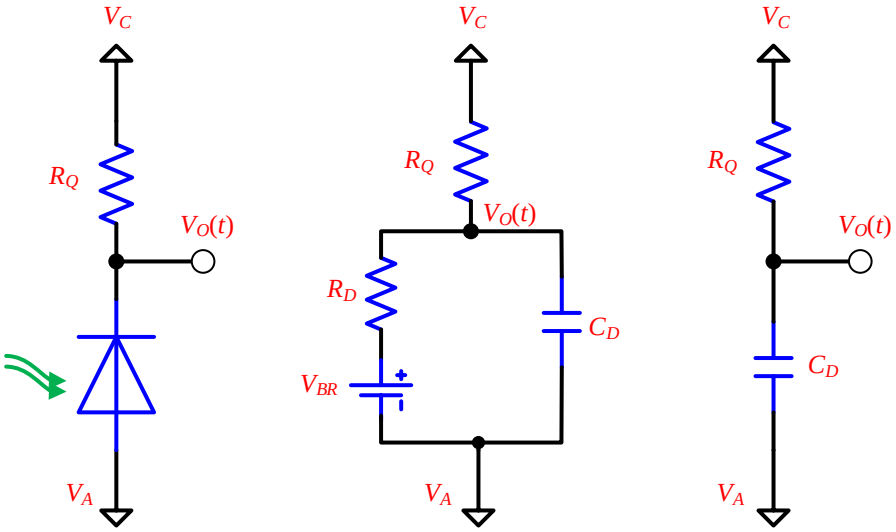


Figure 4.2: (a) SPAD with passive quenching and recharge circuit; (b) Equivalent circuit of passive quenching; (c) Equivalent passive reset circuit

Solving Eq. (4.2), the cathode voltage is

$$V_O(t) = \frac{R_Q(V_A + V_{BR}) + R_D V_C}{R_D + R_Q} + \frac{R_Q V_{ex}}{R_D + R_Q} \exp\left(-\frac{R_D + R_Q}{C_D R_D R_Q} t\right). \tag{4.3}$$

With regards to the recharge process [Figure 4.2(c)], the current function at the cathode is

$$\frac{V_C - V_O(t)}{R_Q} = C_D \frac{\mathrm{d}V_O(t)}{\mathrm{d}t},\ V_O(t=0) = V_C - V_{ex}\,. \tag{4.4}$$

From Eq. (4.4), the cathode voltage is

$$V_O(t) = V_C - V_{ex} \exp\left(-\frac{t}{C_D R_Q}\right). \tag{4.5}$$

From Eq. (4.3) and Eq. (4.5), the time constants for passive quenching and passive recharge can be written as

$$\tau_{Quench} = \frac{C_D R_D R_Q}{R_D + R_Q} = C_D\left(R_D \parallel R_Q\right),\ \tau_{\mathrm{Recharge}} = C_D R_Q\,. \tag{4.6}$$

According to Eq. (4.6), the quench resistor (R_Q) dominates the recharge time, and the quenching time is correlated to the parallel combination of R_Q and R_D. Since R_D is much lower than R_Q, the quenching time is shorter than the recharge time.

Figure 4.3 shows the simulation results of passive quenching and recharge using Eq. (4.3) and Eq. (4.5), from which we can see that the diode is quenched very fast, but it takes a longer time to recharge the diode to its initial state of the cathode potential V_o=V_C. The parameters used in the simulation are listed in Table 4.1.

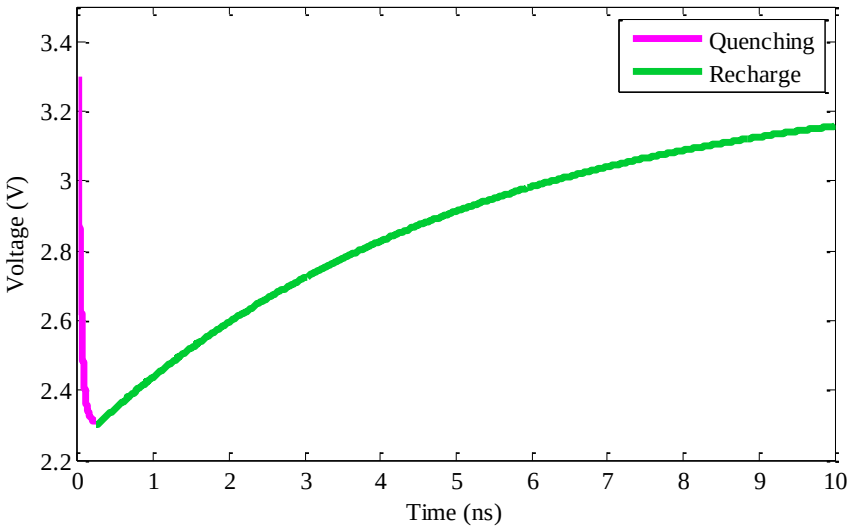


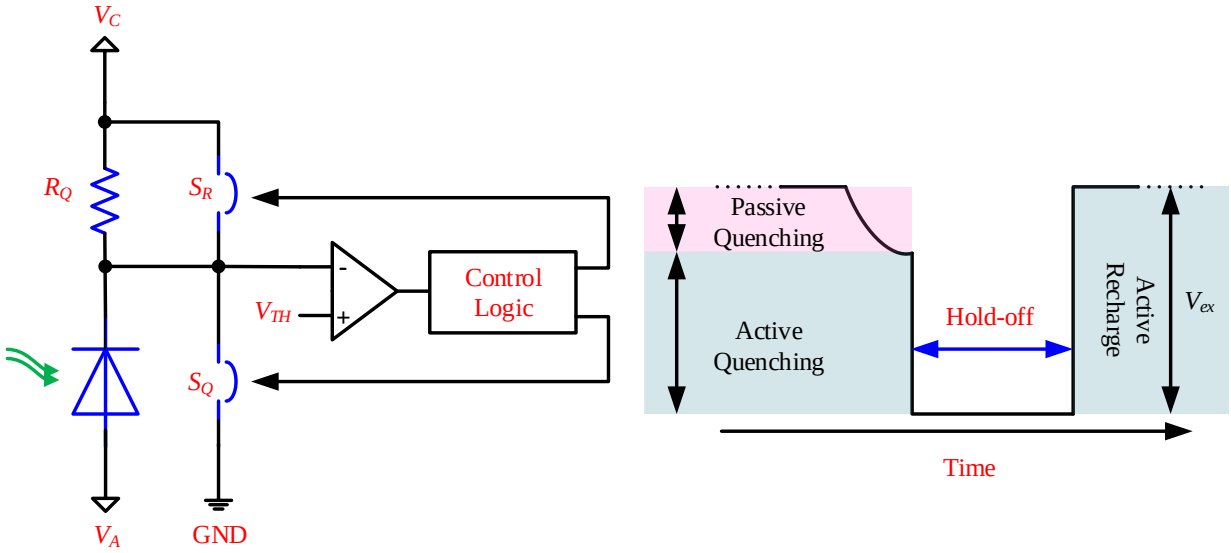
Figure 4.3: Simulation of the SPAD cathode potential $V_o(t)$ with passive quenching and recharge.

Table 4.1: Parameters used for simulation of the passive quenching and recharge circuit

Parameter	Value	Parameter	Value
V_A	-9 V	R_Q	50 kΩ
V_C	3.3 V	R_D	400 Ω
V_{BR}	11.3 V	C_D	100 fF

In a passive quenching circuit, the quench resistor usually occupies an area smaller than the area of the photodiode, so a high fill factor can be achieved. Passive quenching is also suitable for the design of large pixel arrays. A drawback of the passive quenching circuit is the unstable dead time. As depicted in Figure 4.3, the recharge phase is a slow process. If a photon arrives during the recharge phase, since the SPAD is biased at the Geiger mode, this photon could trigger a second detection and stop the recharge process. As a result, the dead time is expanded and longer than the designed dead time.

Another limitation of the passive quenching circuit is the afterpulsing probability. As mentioned above, afterpulsing probability is related to the quenching time and the parasitic capacitance. Since the circuit only contains a resistor and a photodiode, the circuit's capacitance mainly comes from the capacitance of the photodiode. The capacitance of the photodiode is determined by the fabrication process, the SPAD area and reverse bias voltage. For SPADs implemented in deep sub-micron CMOS technologies, the capacitance is small and in the range of fF. This is an advantage of the passive quenching circuit, but the quenching time is long enough for carriers to be trapped. If the SPAD is in the Geiger mode, then trapped carriers can be released later, leading to afterpulsing. In addition, the captured carriers released during the long recharge phase can cause the dead time to increase. To solve the dead time and afterpulsing problems, high-speed control circuits are designed to reduce the quenching and reset times of the SPAD.



(a) (b)

Figure 4.4: [131] (a) Basic diagram of a mixed passive-active quenching circuit; (b) Cathode voltage waveform

Figure 4.4 shows the schematic diagram and the cathode voltage waveform of a mixed passive-active quenching circuit [131]. When an avalanche multiplication is triggered, a large avalanche current flows through the quenching resistor R_Q, which passively quenches the avalanche current initially. The passive quenching reduces the cathode voltage and when this voltage is comparable to the threshold voltage V_{TH} of the comparator, the active circuit is activated to generate a signal through the control logic. The signal from the control logic closes the switch S_Q, actively quenching the avalanche current in a time shorter than with passive quenching. After fully quenched, the SPAD is held-off for a period until a second signal is generated by the control circuit to close the

recharge switch S_R and open the quenching switch S_Q. To quickly recharge the SPAD, the equivalent resistance of the recharge switch is much lower than that of the passive quenching resistor. Upon completing the recharge, the control circuit is deactivated and the SPAD is ready to detect a subsequent photon.

Compared to the passive quenching circuit, the dead time of the SPAD is effectively shortened through the active quenching and recharge. It also mostly solves the problem of dead time expansion problem, since a hold-off of the detector is used, as shown in Figure 4.4(b). Therefore, high counting rates can be achieved for SPADs with active quenching and recharge. As given in [132], a maximum counting rate of 185MHz was achieved with a mixed passive-active quenching circuit, which corresponding to a dead time of 540ps. Regarding the afterpulsing probability, the parasitic capacitance at the node of the SPAD cathode is increased because of the complex connection, but the quenching time is also reduced. Moreover, by adding a hold-off time, the afterpulsing probability is further reduced. For example, the afterpulsing probability given in [132] is only 1.28% with a hold-off time of 5.4ns. However, the fill factor of the mixed quenching circuit is low, since complicated control circuits are used. Hence, SPAD pixels with passive quenching and reset are still the preferred choice for design of large arrays, such as in silicon photomultipliers.

The free running SPADs have been widely used in the time-resolved applications with the TCSPC technique, such as in FLIM and NIRS. Regarding Raman spectroscopy applications, the time-gated SPADs are more suitable, owing to its fluorescence suppression capability.

4.1.2 Time-Gated Operation

The quenching and recharge circuits reviewed in the previous sub-section were mainly developed for free running SPADs [131], [133]. However, there are applications in which the photon detection is required only for a short “time window” after a pulse excitation, and the “time window” has to be precisely synchronized with the excitation. This “windowed” mode of photodetection is known as the time-gated operation of SPAD and also corresponds to the “shutter” in imagers. Typical examples requiring time-gated photodetection include Raman spectroscopy and NIRS. In Raman spectroscopy, time-gated detection is used to remove the strong background fluorescence emitted later than the Raman signal. In NIRS, “early photons” rejection can be achieved by time-gated detection, as introduced in chapter 1. To operate a SPAD in time-gated mode, the control circuit is crucial, because it quickly gates the SPAD above and below the avalanche breakdown voltage and synchronizes the detection with the pulse excitation.

A common way to apply the gating signal is through an AC coupling network, as shown in Figure 4.5(a). A DC pre-bias is provided by a voltage source V_C and the gating signal is coupled through the capacitor C_1. To achieve fast gating, C_1 should be small. However, to maintain long enough pulse duration, C_1 cannot be too small. Thus, the design of the input AC coupling network is important for appropriate gating and sensing of the SPAD.

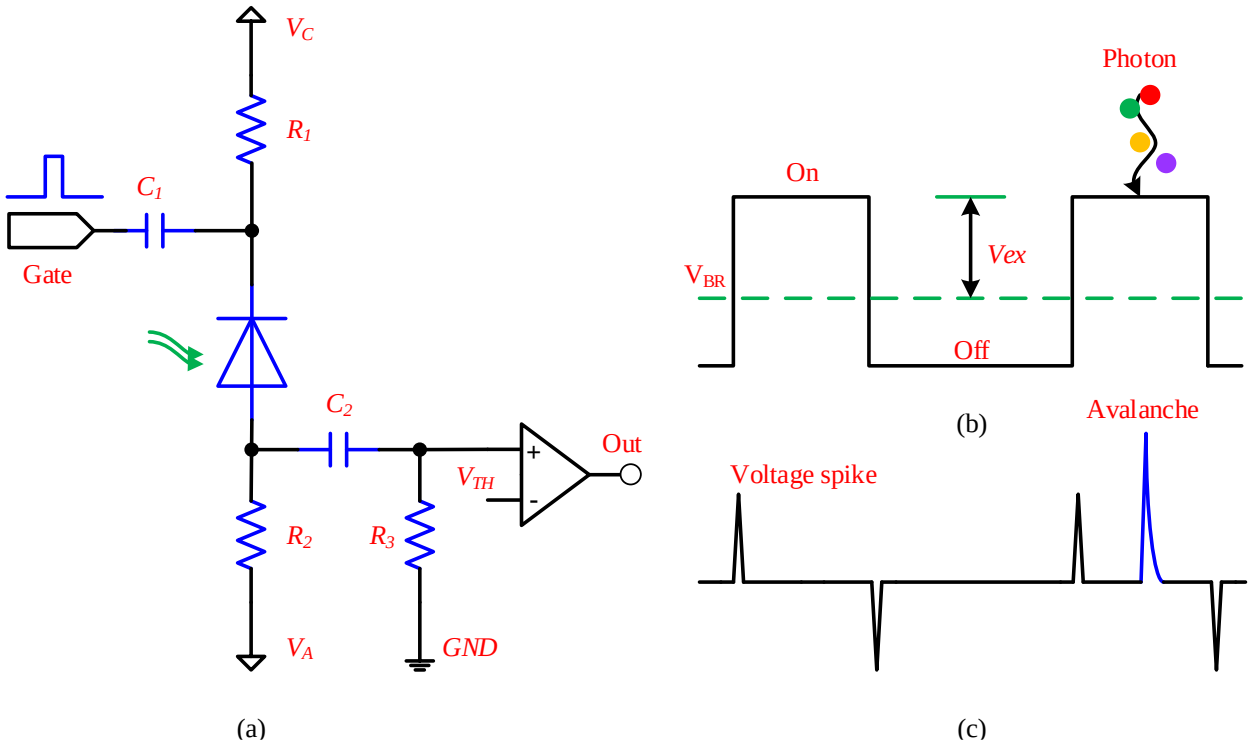


Figure 4.5: (a) Basic diagram of SPAD with time gated control circuit; Timing diagram of the gating signal (b) and the output of the AC pick-up circuit (c)

To sense an avalanche occurring during the narrow gate, usually an AC pick-up circuit is applied and its output is fed to a high speed comparator to extract the detection information. Upon detection, the avalanche current flows through the resistors R_3 ($R_2 \gg R_3$), which quickly quenches the avalanche. Then, the SPAD is brought back to the Geiger mode by R_2. The function of the AC pick-up network is to sense the transient voltage increase of the SPAD's anode that is induced by the avalanche current. However, because of the large R_2, a wide gate window is required, and this limits the SPAD's maximum counting rate.

On the other hand, there are voltage spikes occurring during voltage transitions. As shown in Figure 4.5(b), the SPAD is biased above and below the breakdown voltage. The fast voltage transition introduces a sudden variation to the anode voltage through the voltage divider between the SPAD capacitance and the pick-up network [134]. Sensed by the pick-up network, voltage

spikes (rising and falling) appear at its output, as shown in Figure 4.5(c). To reduce the effect of the spikes, the threshold (V_{TH}) of the comparator should be higher than the amplitude of the voltage spikes. Since the amplitude of the voltage spike is related to the transition speed and amplitude of the gating signal, so the spike amplitude can be higher than the avalanche pulse. Therefore, in this case, setting a higher threshold voltage for the comparator is not efficient. A dummy pixel circuit was used in [134] to solve the spike problem. However, taking into account the large occupied area of capacitors and low fill factor of the circuit shown in Figure 4.5, a more compact time-gated control circuit is designed in this work.

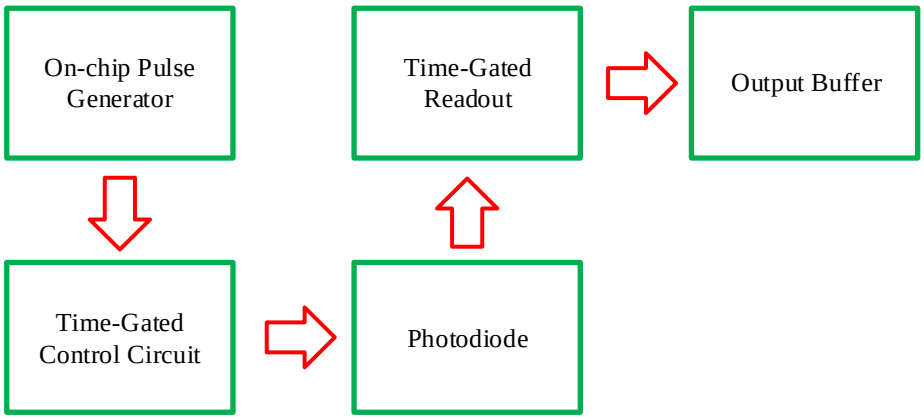


Figure 4.6 Block diagram of the proposed TG-SPAD pixel circuit

4.2 Design of TG-SPAD Pixel Circuit

The block diagram of the proposed TG-SPAD front-end is shown in Figure 4.6. A fast control circuit is designed to control the photodiode, and the sequence of gating signals is provided by on-chip pulse generators that allows for precise synchronization with the excitation pulse. To sense the photon occurrences within the narrow gate window, a time-gated readout circuit is designed. The signal from the readout circuit is conditioned by an output buffer to either drive directly a digital circuit or the load of cables and instruments during the experiments. The details for the components in the block diagram are presented in the following sub-sections.

4.2.1 SPAD Design and Characterization

The design of a photodiode in a standard CMOS technology is restricted by the design rules for the particular technology. There is not much freedom to optimize the layer properties and parameters. Figure 4.7 shows the cross-sectional view of a photodiode designed in the 130nm CMOS technology of IBM. Two diodes exist in this structure, they are the shallow N+/P-well diode and

the P-well/deep N-well diode. The shallow junction diode is used for short wavelength applications, and it has an N-well guard-ring added to prevent the premature edge breakdown. In contrast, the deep junction diode can be used for longer wavelength applications, such as the near infrared region. The optical response of these two diodes are provided later in this section. Since the target wavelength band of our application is between 546nm and 633nm, then the shallow junction diode is used, but the deep junction diode was also designed and characterized, since the breakdown voltage of the deep junction limits the maximum voltage that can be applied to the anode of the shallow junction.

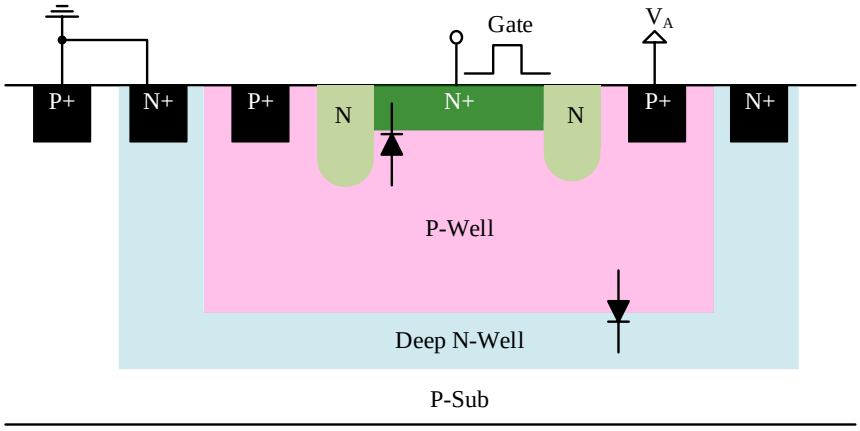
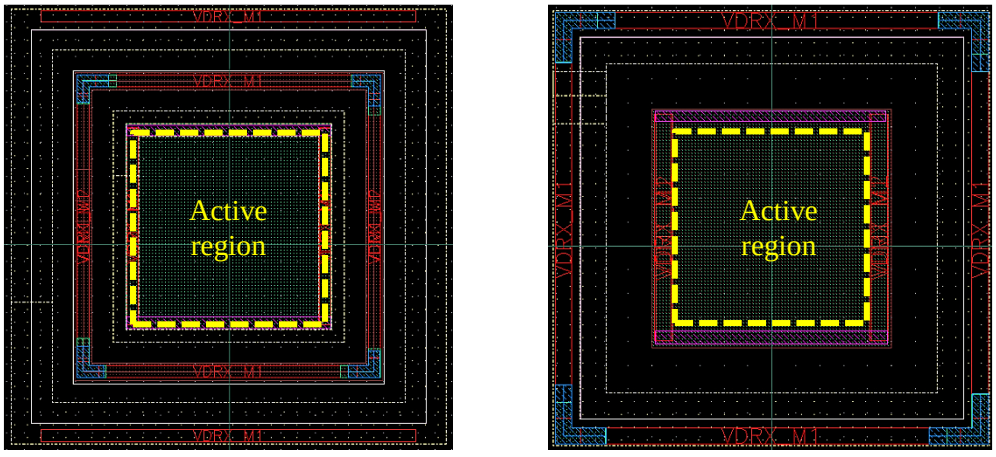


Figure 4.7 Cross section view of photodiode implemented in CMOS technology



(a) (b)

Figure 4.8: Layout top view of the N+/P-well diode (a) and P-well/Deep N-well diode (b)

Figure 4.8 shows the top view of the layouts of the N+/P-well diode and the P-well/Deep N-well diode. The square-shaped active regions of these two diodes are 10 x 10µm^2, surrounded by space for accessing of the layers in the SPAD structure. While a circular structure is preferred to eliminate sharp junction corners or edges that cause premature edge breakdown, a circular shape

is not allowed in standard CMOS technology. In addition, due to the technology rules for minimum spacing and widths, the photo-active area of the SPAD is only ~16% of the total area of the SPAD layout design. Hence, the fill factor of the TG-SPAD pixel will be less than 16%, when the control circuits are added.

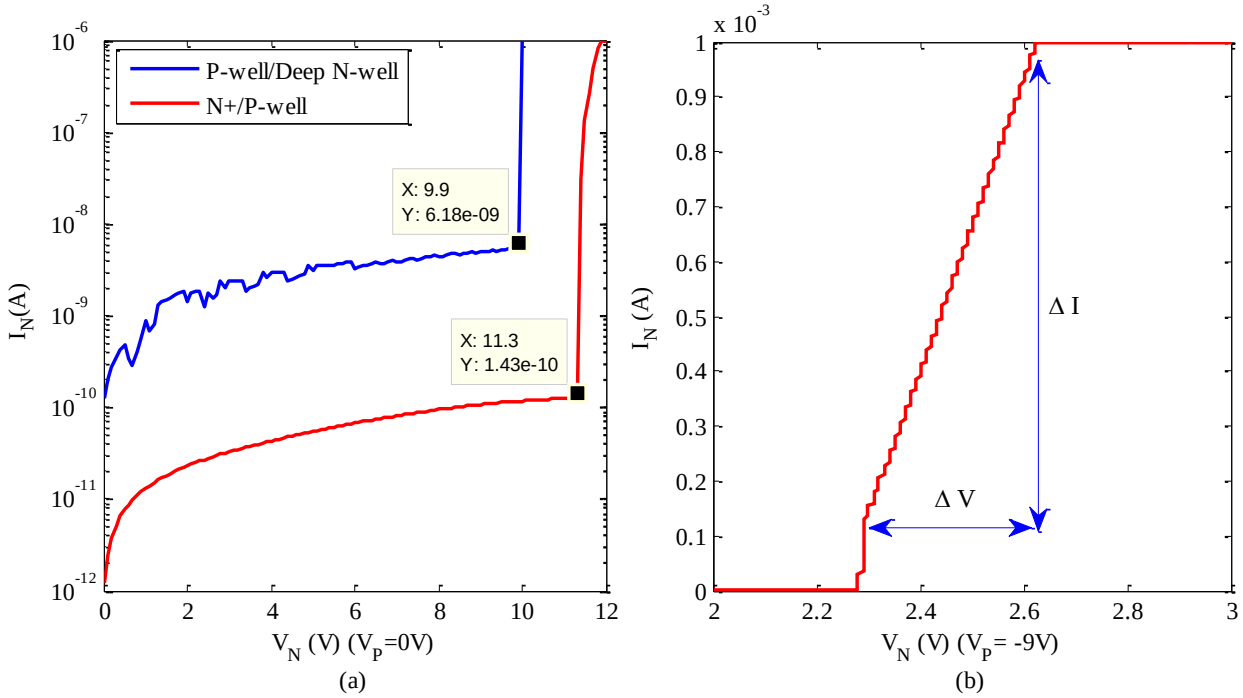


Figure 4.9: (a) I-V measurements of the two diodes with P region grounded; (b) Extraction of the dynamic resistance of the N+/P-well diode

The IV characteristics of the two diodes were measured on-chip in a probe station. The DC bias was provided by the Agilent Semiconductor Parameter Analyzer 4156C, which also measures the nodal current. Figure 4.9(a) shows the I-V curves of the two diodes. The breakdown points are indicated by labelled squares on the curves. The avalanche breakdown voltages of the N+/P-well diode and the P-well/Deep N-well diode are 11.3V and 9.9V respectively.

As mentioned in section 4.1.1, the dynamic resistance of the photodiode is important when determining the quenching speed. To measure the dynamic resistance, the I-V curve of the N+/P-well diode was re-measured around the breakdown, biasing the P-well (anode) at V_A=-9V, and sweeping the bias of the N+ region (cathode) from V_C=0V to 3V. The current compliance was 1mA, as indicated by the horizontal portion in the I-V curve on the top of Figure 4.9(b). This biasing conditioning of the measurement were selected to reflect the operation in the time-gated mode. In this operation, the P-well is negatively biased, and the N+ region is switched between *GND* and

V_{DD} (3.3V). The portion of the characteristic indicated with ΔI and ΔV is shown in Figure 4.9(b), and the dynamic resistance (R_D) of the SPAD in the avalanche breakdown region is determined by

$$R_D = \frac{\Delta V}{\Delta I} \approx \frac{0.32\text{V}}{0.83\text{mA}} = 386\Omega \tag{4.7}$$

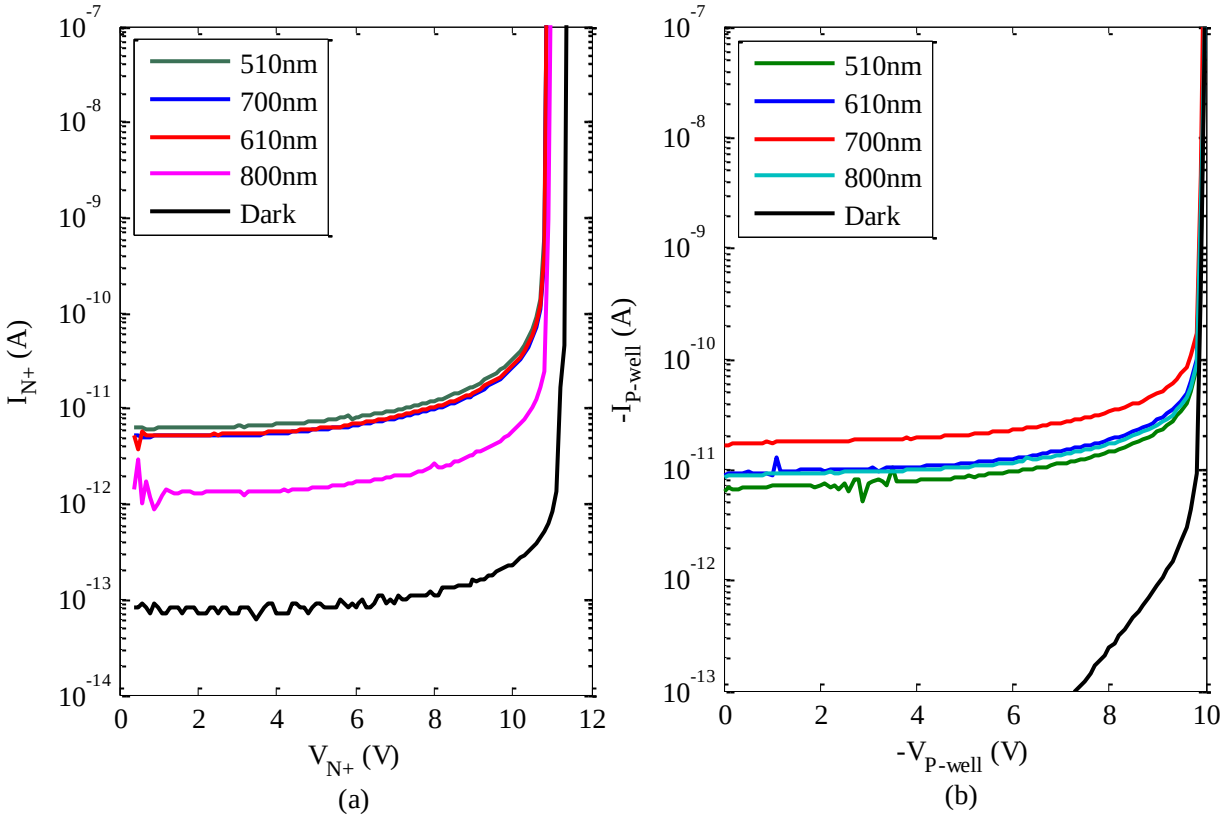


Figure 4.10: Optical response measurements: (a) N+/P-well diode; (b) P-well/Deep N-well diode

The space charge region capacitor C_D is also very important for a SPAD operated in the time-gated mode, since sub-nanosecond gating signal is sometimes applied. The capacitance was measured to be ~100fF. C_D and R_D determine the quenching time, by the time constant $\tau_{Quench} \approx C_D R_D$, as given in Eq. (4.6).

The optical response of the two diodes was measured. The applied light source was a lamp, filtered by optical filters of 10nm bandwidth. Figure 4.10 shows the measured photocurrent from visible to near infrared regions (fixed photon flux: ~3×10^8 photons/s). The diodes were also measured in dark as a reference. Comparing the optical responses of the two diodes, the shallow junction diode [Figure 4.10(a)] has higher photo sensitivity in visible region. However, when the incident wavelength goes up to 800nm, the photon current drops very fast, close to the current

measured in dark. In contrast, the deep junction diode [Figure 4.10(b)] has higher sensitivity in the near infrared region (~700nm), making it suitable for NIRS applications.

4.2.2 On-chip Pulse Generator Design and Characterization

For the purpose of fast gating the SPAD and simple synchronization of the gate window, on-chip pulse generation is critical [135]-[137]. The pulse width and its stability are important for the time-gated operation of the photodiode.

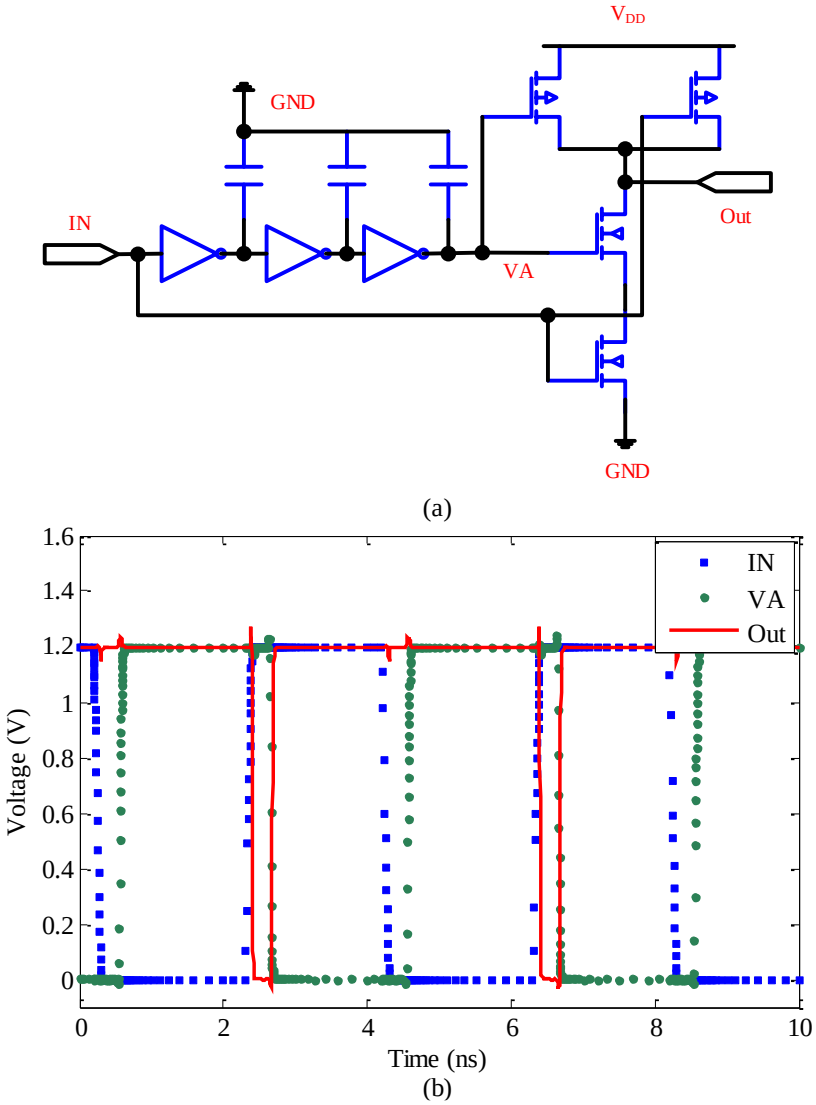


Figure 4.11: (a) Circuit topology of the on-chip pulse generator; (b) Waveform in the pulse generator by Cadence

Figure 4.11(a) shows the schematic of the on-chip pulse generator, which consists of a delay chain and a NAND gate. Figure 4.11(b) shows the timing diagram of the generator, simulated by Cadence Virtuoso. A transition of IN from low to high signifies the beginning of the pulse

generation. This transition causes Out=low, since VA is still high. The input signal is delayed by the delay chain, and inverted to VA=low after the delay, which bring back Out=high, and completed the generation of the output pulse with active level Out=low. Thus, the pulse generator produces a negative pulse at the rising edge of the input signal, and the pulse width is equal to the delay in the delay chain. The delay can be adjusted either by choosing different odd numbers of delay units or by adjusting the load capacitance C in each delay unit, because the delay of the delay unit is determined by its RC time constant.

As shown in Figure 4.11(b), a narrow negative pulse (~280ps) is generated at the rising edge of the input signal. VA is a delayed and inverted version of IN, and the pulse width is equal to the delay between IN and VA. To test the efficiency of this circuit in generating narrow pulse, a narrow pulse width was designed and tested, as follows.

Chips of the pulse generator were fabricated in the 130nm CMOS technology and mounted on a probe station for on-chip testing. DC supply was provided by the Agilent Semiconductor Parameter Analyzer 4156C. The input clock signal was provided by a pulse pattern generator (MP1763B, Anritsu). Its bit rate was set by the frequency of a Synthesized Sweeper (83752A, Agilent). The output Out was measured by the Agilent wide-bandwidth oscilloscope 86100A.

Before measurement, the experimental setup was calibrated using a 200μm length and 50Ω calibration substrate. The purpose of this calibration was to check the connections between the RF cables and all the instruments. The input was a clock signal with 250MHz frequency and 1.2V amplitude. To protect the oscilloscope from burning, RF attenuators were connected at the oscilloscope inputs, bringing all signals in the range between 0.2V and 0.5V.

Figure 4.12(a) shows the calibration result and the pulses measured from six chips. The curve obtained from the calibration test proves for the correct connections of the setup. Stable and narrow pulses were generated based on the measurements of six chips. The pulses were always generated at the rising edge of the input clock signal, as designed. Considering the attenuation of the 6dB attenuator, the measured amplitude of the pulses is ~0.6V.

Because the sensor is designed for field applications, the environmental temperature variations were also considered. Figure 4.12(b) shows the measured results with temperatures from 50°C to 170°C. Both the pulse width and amplitude were measured to be stable at different temperatures, which proved that the on-chip generator functions very well at high temperatures.

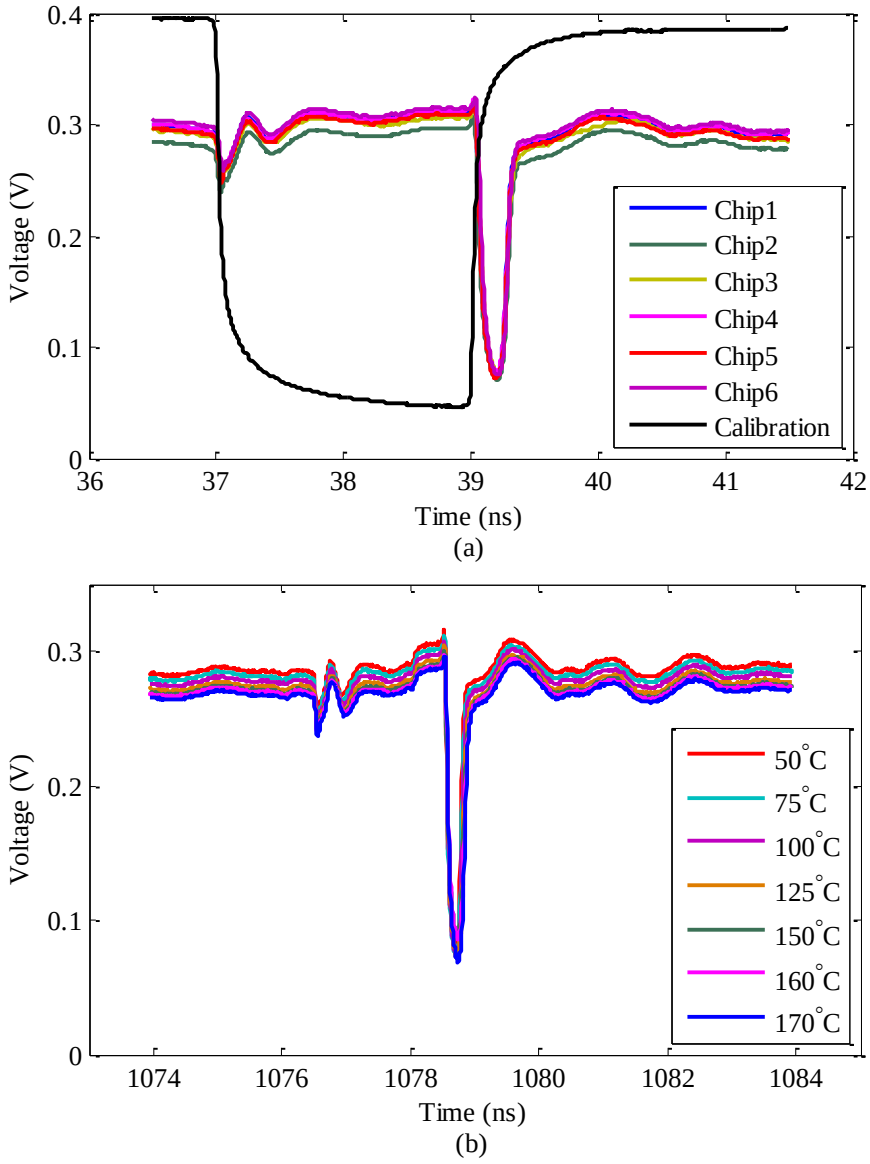


Figure 4.12: (a) Pulses measured from six chips; (b) Pulse measured with different temperatures

4.2.3 Design of TG-SPAD Pixel Circuit

After characterization of the photodiode and pulse generator, the TG-SPAD front-end is designed. From Figure 4.6, the circuits to be designed are the control and time-gated readout circuits. To operate a SPAD in the time-gated mode, the control circuit must be capable of gating the SPAD in a very short gate window, which implies that fast quenching and recharge of the SPAD is required. In addition, since the final goal is to design an array, the pixel circuit cannot be very complex, in order to preserve a reasonable fill factor.

As discussed in section 4.1, the large quenching resistor is the cause for the slow recharge process. While able to provide fast gating of the SPAD, the circuit topology shown in Figure 4.5(a)

is also not very suitable, because the AC coupling capacitors would occupy a large area on chip. Thus, the design of time-gated SPAD pixel should use active reset and include all other circuits for control and signal conditioning, avoiding AC coupled gating. In short, the circuit must be compact and simple.

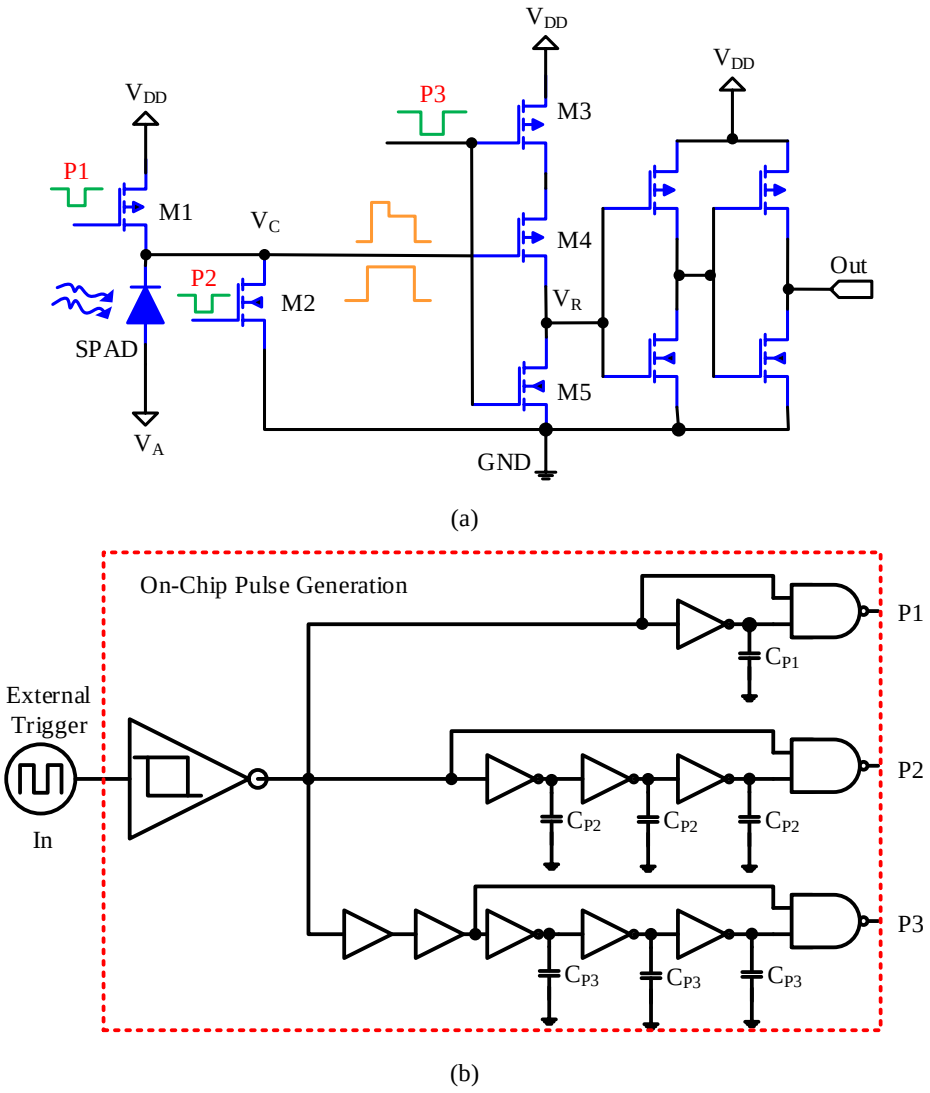


Figure 4.13: Schematics of the (a) proposed TG-SPAD pixel circuit, (b) on-chip gating control circuits

Figure 4.13(a) shows the schematic of the proposed TG-SPAD pixel circuit, including the control, time-gated readout, and buffer circuits. If the buffer circuit is not counted, then the front-end consists of only 5 transistors, which helps in improving the pixel's fill factor. The fast gating of the photodiode is realized by two transistors, M1 and M2. M1 pulls the cathode voltage up to V_{DD}, and M2 is used to connect the cathode to GND. The time-gated readout is realized by three series connected transistors, M3, M4, and M5. The buffer circuit consists of a cascade of two

inverters, forming digital levels at the output. A sequence of three pulses P1, P2, and P3 is used to drive the pixel circuit, which provides reset, time-gated detection and readout upon the rising edge of the external trigger signal. A schematic diagram of the generation circuits for pulses P1, P2, and P3 is shown in Figure 4.13(b), on basis of the pulse generator designed in section 4.2.2. The Schmitt trigger is used to condition the external trigger signal to the target rectangular shape and amplitude (3.3V) before being applied for pulse generation. This signal conditioning of the external triggering signal is essential for the proper operation of the on-chip pulse generators.

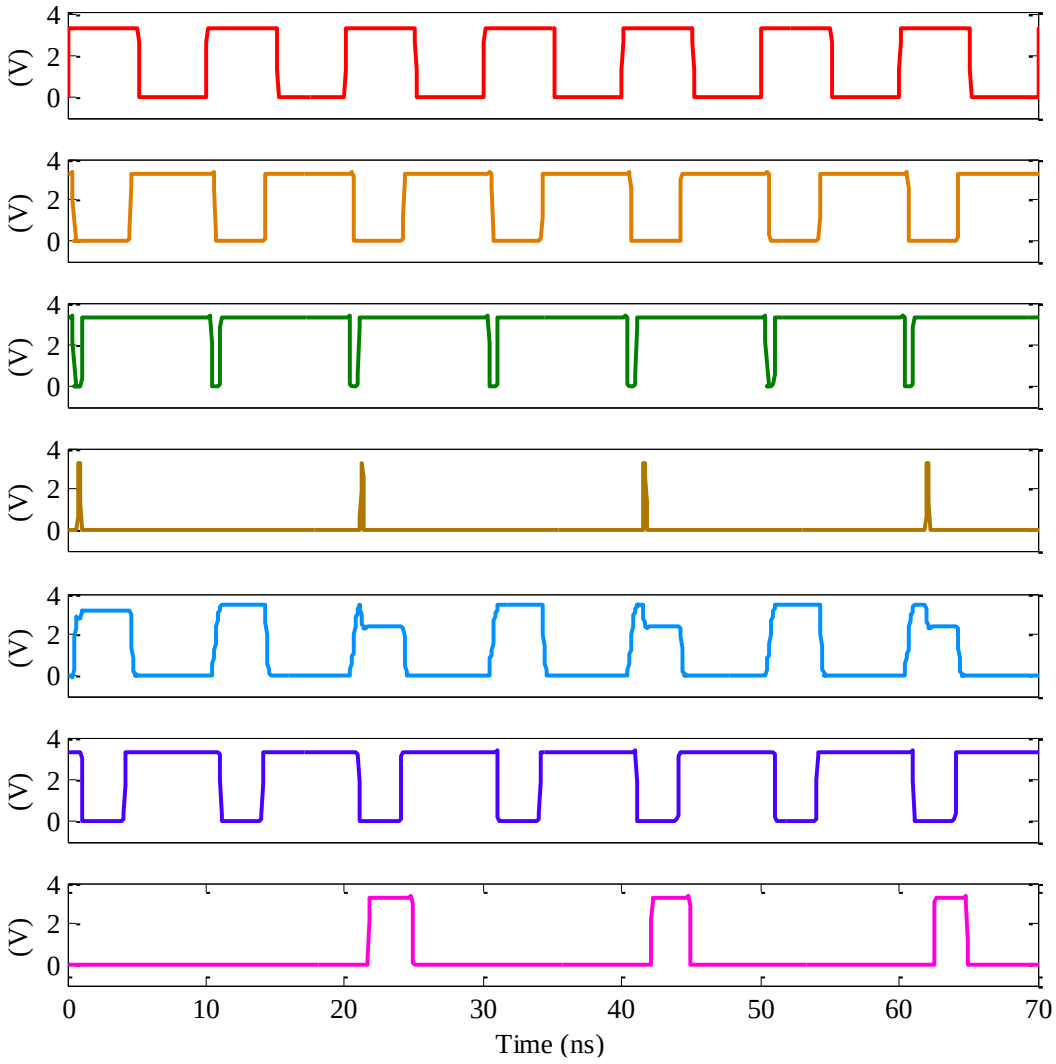


Figure 4.14: Timing diagram of the TG-SPAD pixel circuit (Simulated by Cadence Virtuoso)

Figure 4.14 shows the timing diagram of the TG-SPAD pixel circuit, simulated by Cadence Virtuoso. Each gate window starts at the falling edges of P1 and P2, which simultaneously turn on M1 and turn off M2. M2 isolates the cathode from GND, and M1 pulls the cathode voltage to V_{DD}, charging the SPAD to (V_{BR}+V_{ex}). The pulse width of P1 is very short (~500ps), which means that

the reset time (t_{RST} = 0.5ns) of the SPAD is also very short. After the short reset, P1 goes high, M1 is turned off and the SPAD's cathode is then isolated from both GND and V_{DD}. This transition signifies the onset of the gate window. The gate window continues when both M1 and M2 are off. The SPAD's capacitance maintains the high cathode voltage $V_C \approx V_{DD}$ of the reset state during the gate window, and a photon detection event will cause a step discharge, as detailed shortly.

The pulse P2 is longer (~4ns) than the reset pulse P1, and M2 is in the off state until the rising edge of P2. Therefore, the overall gate window (t_{win}=3.5ns) of the SPAD is the time interval between the rising edges of P1 and P2, which is the same as the difference of the duration of the low levels of P2 and P1.

Unlike the free running mode, transistor M1 in the TG-SPAD is in the off state during the rest of the gate window, so the cathode voltage is not recharged back to V_{DD}, but held off, until the start of next gate window. Thus, the hold-off time, t_{off}, of the TG-SPAD is defined in Eq. (4.8), where f_G is the gating frequency.

$$t_{off} = 1/f_G - (t_{win} + t_{RST}) \tag{4.8}$$

To simulate the avalanche process, an electric pulse was used to mimic a photon signal. Since the avalanche multiplication occurs in a very short time, the pulse width of the photon signal was set to 100ps. When a detection event occurs within the gate window, electron-hole pairs are generated and multiplied in the high-field depletion region of the SPAD. A large avalanche current discharges the SPAD (and the other capacitances connected to the cathode) down to the SPAD breakdown voltage. A significant down-step of the cathode voltage (with amplitude of V_{ex}) is present at the moment of the photon arrival, as shown in Figure 4.14. However, if a photon arrives before the gate window, that is when M1 is turned on, the relatively small equivalent resistor of M1 is not able to fully quench the SPAD (see the first photon shown in Figure 4.14).

A time-gated readout circuit is designed to sense the cathode voltage drop only within the gate window. This readout circuit consists of three series-connected transistors (M3, M4 and M5), triggered by the pulse P3. Initially, the high level of P3 turns off M3 and connects V_R to GND. The readout starts at the falling edge of P3, immediately after the reset has finished at the rising edge of P1. The low level of P3 turns on M3 to provide supply for M4 and turns off M5 to disconnect node V_R from GND. If no detection occurs within the gate window, the high level of V_C keeps M4 off, and V_R is maintained to be zero. In contrast, if a detection event occurs within the gate window, the voltage drop of V_C will turn M4 on to pull up V_R from GND to a high level (V_{DD}-2V_{TH}). The

readout ends at the rising edge of P3, resetting V_R to zero. The unlabeled transistors at the output are two inverters that form the levels between V_{DD} and GND and provide for signal buffering.

Note that if no photon is detected, there is no pulse at the pixel output (terminal Out). Thus, the time-gated SPAD can be used for photon counting. In addition, it is worth to note that the pulse duration of the signal at Out is equal to the time interval between the photon arrival and the end of the gate window. As shown in Figure 4.14, the pulse width at Out varies with the different photon arrival times. Thus, this TG-SPAD pixel can be used to additionally extract the information about photon arrival time within each gate window.

To count the number of photons detected in a given period, either digital or analog counters can be designed. A digital counter usually requires more transistors than an analog counter [133]. Considering the area usage, the analog counter is more suitable for in-pixel design. To be compatible with the analog counter, the variable pulse obtained at the output of the circuit in Figure 4.13(a) can be conditioned to a fixed width and amplitude by an on-chip wave shaping circuit.

Figure 4.15(a) shows the schematic diagram of the wave shaping circuit, which consists of a RS flip flop and a delay unit. Assume that the signal to be shaped (IN) is the output signal of the circuit in Figure 4.13(a). Initially, IN is low, since Qb is high, the inverter keeps Reset at low. When IN goes to high, it sets the output Out of the RS flip flop to high and the complementary node Qb to low. Since the delay in the inverter keeps Reset also low for the time of the delay, the output Out is high during the inverter delay, and the RS flip flop is locked in this state even if the short IN goes low during the inverter delay. After the inverter delay, Reset goes high, Out goes low and the circuit recovers the initial state. Thus, as illustrated in Figure 4.15(b), the wave shaping circuit produces constant-width pulse at Out at any shorter-width pulse of IN. The delay in the inverter of the wave shaping circuit was chosen to be slightly longer than the gate window (t_{win}=3.5ns), so that the pulse width ~3.9ns is present of the output of the wave shaping circuit. The adjustment of the delay was made during the simulation by adjusting the sizes of transistors M6 and M7 and the capacitance C_S.

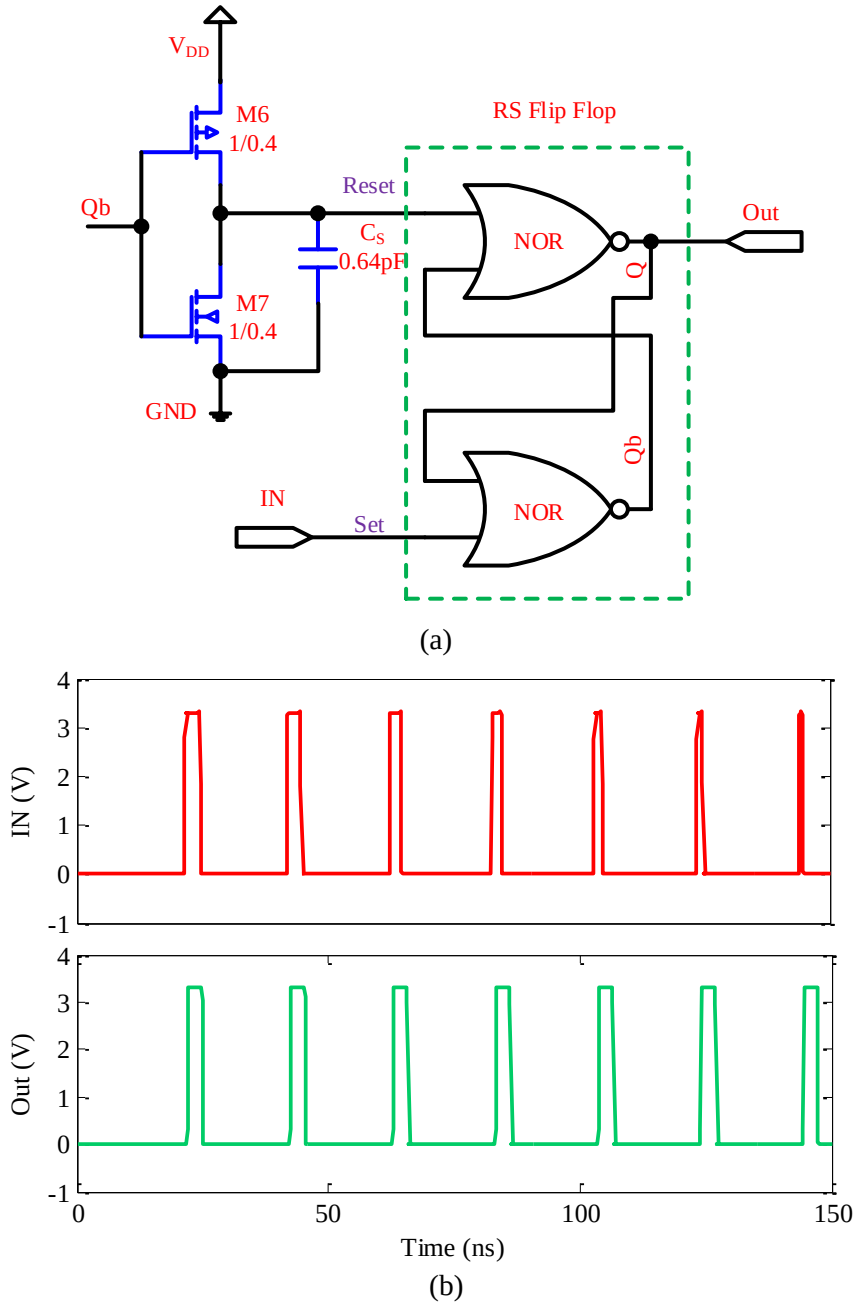


Figure 4.15: (a) Schematic diagram and (b) timing diagram of the wave shaping circuit

The design of the time-gated SPAD front-end was laid out on a 1mm x 1mm chip. The layout view is shown in Figure 4.16. This chip contains two pixels (Area 2). One pixel circuit is with the wave shaping circuit, and the other is without the wave shaping circuit. The on-chip gating circuit is depicted as Area 3, and the large dashed block (Area 1) is a test structure. Fill factor of the TG-SPAD pixel is ~9.8% and as discussed in section 4.2.1, this is restricted by the design rules of this particular CMOS process.

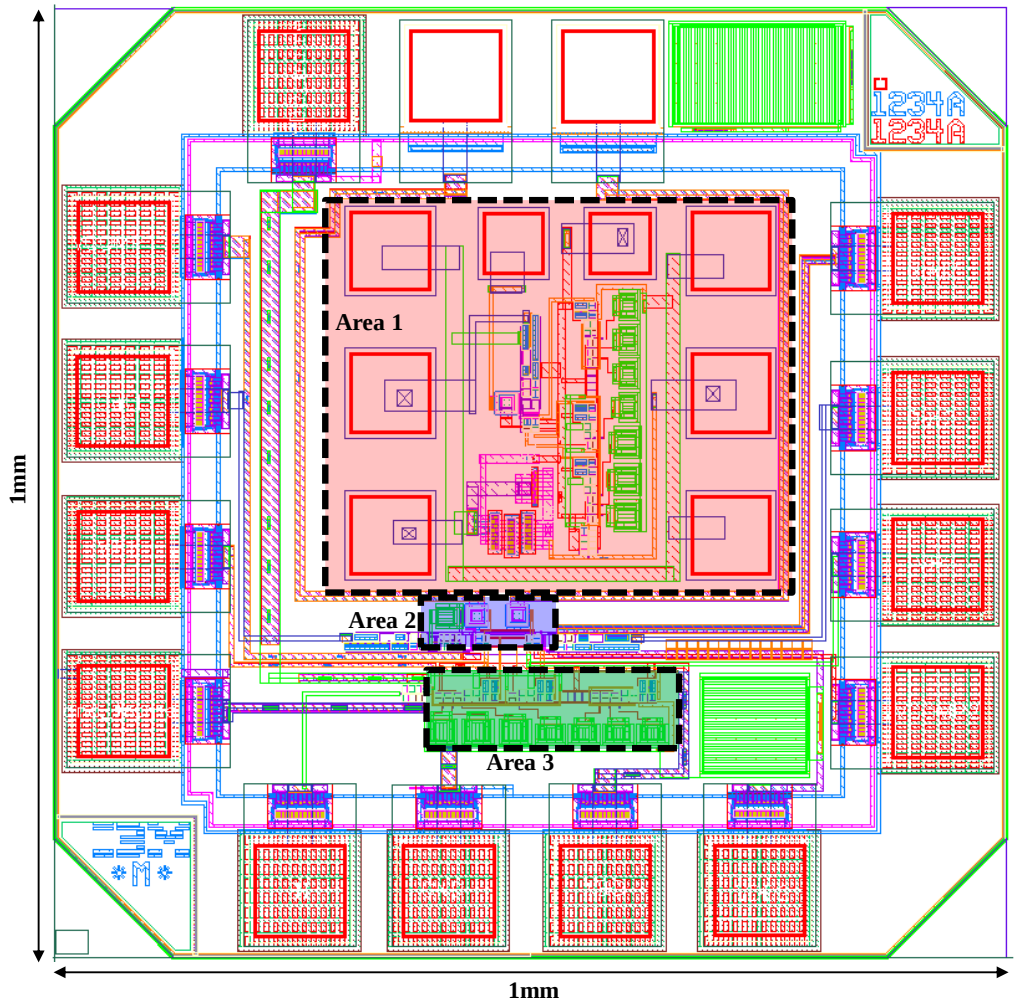
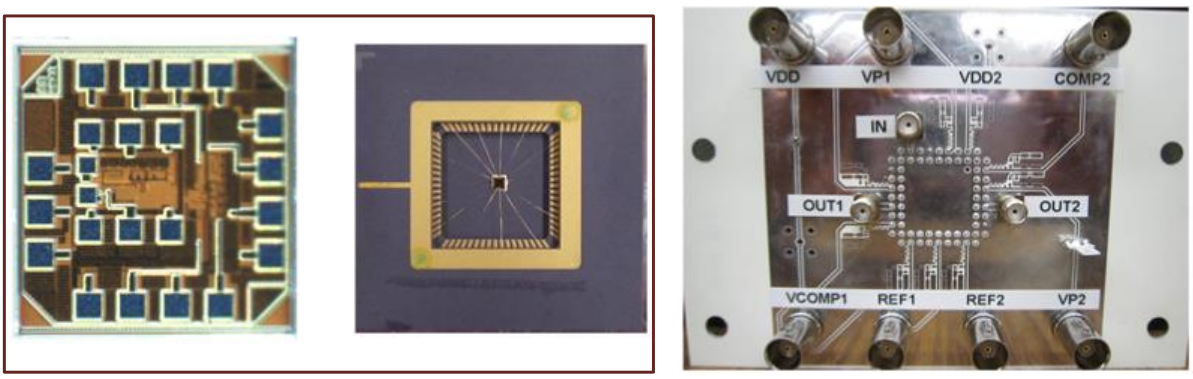


Figure 4.16: Layout view of the designed chip with the prototype of the time-gated SPAD front-end

4.3 Measurements of TG-SPAD Pixel Circuit

Chips with the prototypes of the designed TG-SPAD front-end were fabricated in the 130nm CMOS technology of IBM. The packaged chip was mounted on a custom printed circuit board (PCB). Figure 4.17(a) shows the micrograph of the fabricated TG-SPAD and the image of the packaged chip. A photograph of the designed PCB is shown in Figure 4.17(b), with labels of the DC and RF connectors. The chip is mounted on the other side of the PCB. To minimize the effect of spikes and other fluctuations in the DC bias supply, multiple by-pass capacitors were added between the bias lines and ground. The prototypes are evaluated for functionality and performance, as discussed in the next sub-sections.



(a) (b)

Figure 4.17: (a) Micrograph of the fabricated TG-SPAD and the packaged chip; (b) PCB with DC and RF connectors and by-pass capacitors

4.3.1 Functional Test of the TG-SPAD Pixel Circuit

The chips were tested to determine their functionality. Bias voltages were provided by the Agilent B1500A Semiconductor Device Analyzer (SDA), which also monitored the bias current during the experiments. The input trigger signal was provided with a pulse pattern generator (MP1763B, Anritsu). Its bit rate was set by the frequency of a Synthesized Sweeper (83752A, Agilent). The output Out was measured with a high-speed real-time oscilloscope (LeCroy SDA 18000 Serial Data Analyzer). To test the photon detection capability of the prototype, the chip was illuminated with a halogen lamp, to obtain random photon arrival time. Thus, positive pulses of random durations are expected to be present at the output of the TG-SPAD front-end without wave shaping in-pixel circuit. In addition, the intensity of the lamp was adjusted so as not to saturate the SPAD during the measurement. Therefore, the waveform of the font-end output would contain both situations of photon detection and no photon detection.

The chip was operated at a gating frequency of 50MHz, which corresponds to a cycle of 20ns. From the design of on-chip pulse generators, the gate window is expected to be 3.5ns. The pulses measured at the output of the TG-SPAD front-end are shown in Figure 4.18(b). The time span of this signal acquisition is 100ns, containing 5 gate windows in total. Among these 5 gate windows, photon detections occurred in 4. No photon was detected in the third gate window in Figure 4.18(a), and the output was zero. In addition, as expected from the design, variable pulse widths are present, when the photons arrive randomly. This is determined by the different photon arrival times.

Due to the impedance mismatch between the chip and test setup, the maximum amplitude of the Out signal drops from 3.3V (VDD) to ~2.2V. Moreover, the parasitic capacitor and inductor

from the chip package and PCB cause the rise time of the Out signal to increase to hundreds of picoseconds, which further reduces the amplitude of narrower signals, as measured in Figure 4.18(b). The simulation results considering RF reflection and parasitics are given in Figure 4.18(a), which confirms the reduction of the amplitude due to impedance mismatch and parasitics in the test set-up.

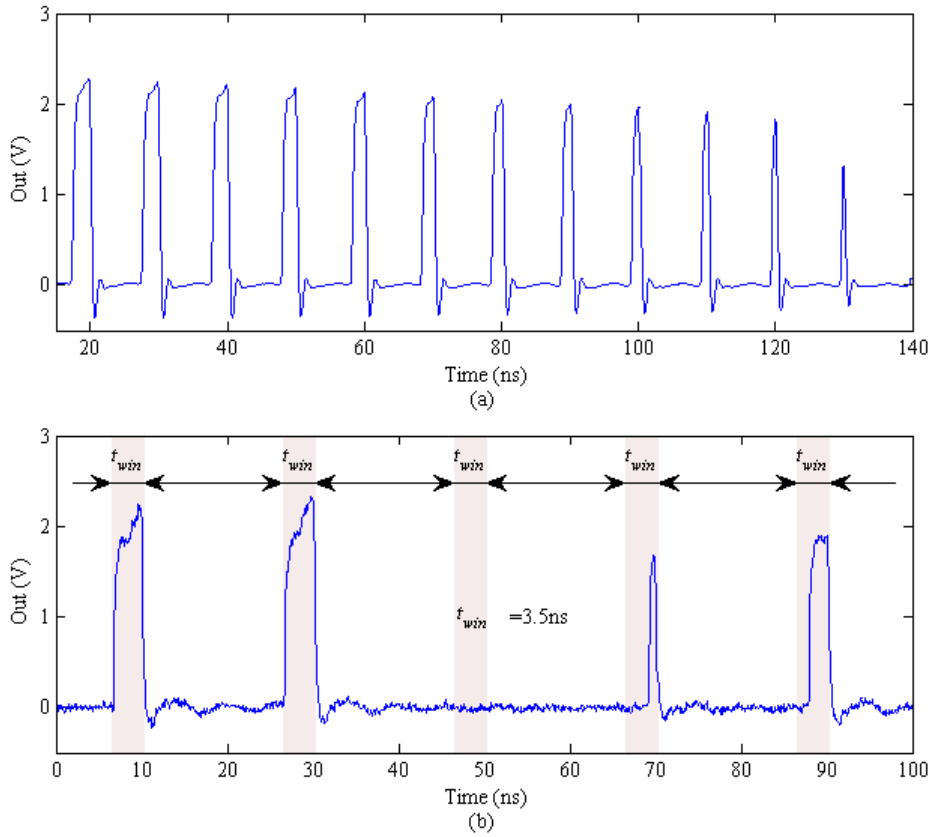


Figure 4.18: Outputs of the TG-SPAD pixel circuit: (a) Simulation; (b) Measurement

The measured result of the pixel with wave shaping circuit is shown in Figure 4.19. The time span of this measurement is 200ns, corresponding to 10 gate windows in total. Five pulses were obtained at the output of the pixel circuit. In contrast to the results shown in Figure 4.18(b), pulses obtained in this pixel circuit were fixed to a constant width and height. Combined with an analog counter, the number of photons detected in a given period can be counted.

As mentioned in section 4.2.3, the pulse width of the TG-SPAD output is equal to the time interval between the photon arrival and the end of the gate window. Thus, photons arriving at the beginning of the gate window can generate wider pulses than photons arriving at the end of the gate window. Figure 4.20(a) presents the histogram of the pulse width. Owing to the random arrival of photon, a wide distribution of the pulse width is present. Observe that the histogram is for the

maximum width of 3.54ns, and no pulse width longer than 3.54ns was detected. This confirms the 3.5ns gate window, as designed.

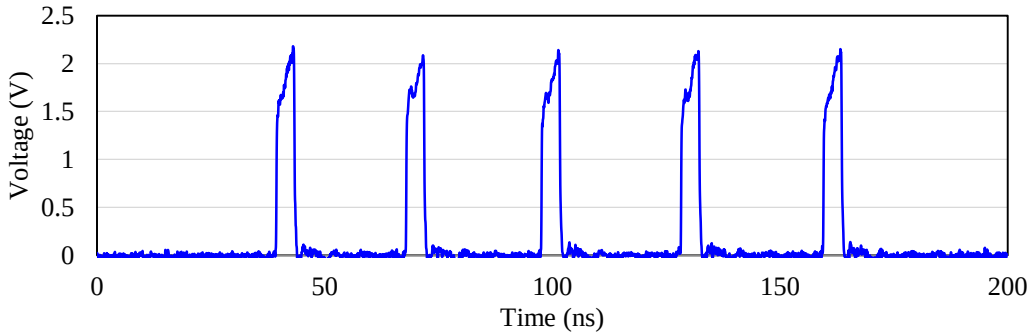


Figure 4.19: Output of TG-SPAD pixel circuit with in-pixel wave shaping circuit

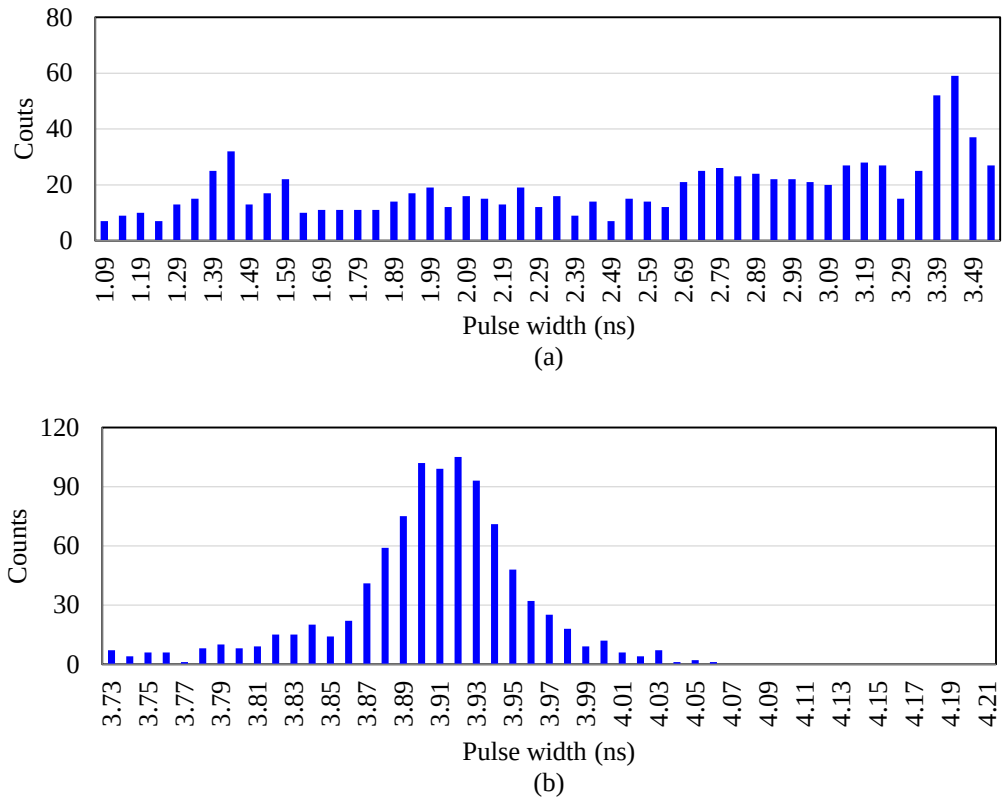


Figure 4.20: Histogram of the output pulse width: (a) pixel without wave shaping; (b) pixel with wave shaping

In this design, from all photons arriving during the 3.5ns gate window, only the first photon will be detected. In the particular test with strong illumination, most of the photon detections will occur during the early time of the gate windows. Figure 4.20(b) shows the histogram of the pulse width formatted by the wave shaping circuit, from which we can see that the pulse width is shaped to ~3.9ns. The FWHM of the distribution of the widths of the formatted pulses is less than 100ps.

Overall, the functional test of the chip under illumination confirmed that the operation of the chip is as expected from the design, detecting the random arrival of photons in a 3.5ns gate window and producing signals with virtually constant width with the wave shaping circuit.

4.3.2 Performance Tests of TG-SPAD Pixel Circuit

Measurements in section 4.3.1 have proved that the TG-SPAD functions as designed. Before being used in applications for photon detection, the performance of the TG-SPAD is determined. Important performance parameters of a TG-SPAD are the dark count probability and photon detection efficiency.

4.3.2.1 Dark Count Probability per Gate Window (DCP_{GW})

Dark count is a key performance parameter of a SPAD. It determines the minimum photon rate that can be distinguished. For TG-SPADs, photons are detected by gate windows, so dark count probability per gate window (DCP_{GW}) is usually evaluated. The effective dark count rate (DCR) of this TG-SPAD front-end can be re-calculated from DCP_{GW} according to Eq. (4.9).

$$\mathrm{DCR} = \mathrm{DCP_{GW}} / t_{win} = \mathrm{DCP_{GW}} / 3.5ns\,. \tag{4.9}$$

A major source of dark counts in CMOS SPADs at room temperature is the thermally generated carriers. Generation-recombination (GR) centers, introduced into the energy band-gap by defects, can trap and release carriers according to Shockley-Read-Hall (SRH) statistics. Release of trapped carriers trigger dark counts [138]. For SPADs fabricated by deep sub-micron (DSM) CMOS technologies, the increased doping level together with the narrowed depletion region, enhance tunneling effect — band-to-band or trap-assisted [139]. Therefore, tunneling in DSM CMOS SPADs becomes another significant source of dark counts. Thus, SPADs implemented in DSM CMOS technologies feature high dark counts and afterpulsing probabilities.

Thermal generation is strongly temperature dependent [140], [141]. Considering only thermal generation, the temperature dependence of DCP_{GW} [66] can be written as

$$\mathrm{DCP_{GW}} \propto T^2 \exp\left(-\frac{E_g}{2kT}\right) \tag{4.10}$$

Here, k is the Boltzmann constant, T is the absolute temperature, and E_g is the bandgap energy. The slopes in Arrhenius plots ($\ln(DCP_{GW}/T^2)$*vs. 1/kT*) are the thermal activation energy E_a of generation, and $E_a \approx ½E_g$ should be approximately half of the band gap energy in doped semiconductor, because mid-gap defects are the most efficient GR centers. However, it is worth noting that for good quality

SPADs, thermal generation of initial carriers in a depleted semiconductor can be due to band-to-band generation-recombination and diffusion of minority carriers from contact regions [142], [143], having activation energy equal to the band gap of silicon (E_g~1.1eV). Thus, activation energy is an indicator of the origin of dark counts and the quality of the fabrication process.

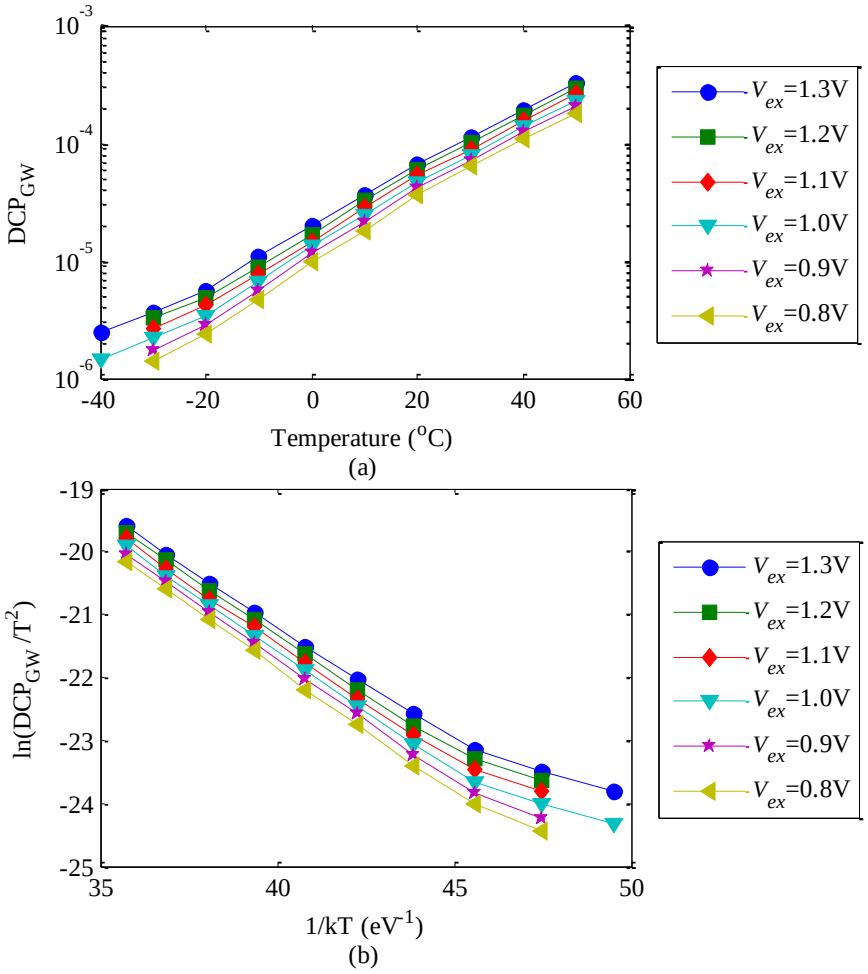


Figure 4.21: (a) Temperature and excess bias dependence of dark count probability per gate window (DCP_{GW}) of the TG-SPAD front-end; (b) Arrhenius plot of DCP_{GW}

To identify the origin of dark counts in this SPAD, DCP_{GW} has been measured over a temperature range from −40°C to +50°C and at different excess bias voltages V_{ex}. Fig. 4.21(a) presents results from the temperature measurements at six bias levels. In Fig. 4.21(a), DCP_{GW} increases with temperature, which implies a thermal activation of DCP_{GW}. Fig. 4.21(b) shows the Arrhenius plot of the data, from which we can see that E_a is temperature dependent, since the slopes in the Arrhenius plot are different at low and high temperatures. The values of E_a, as extracted from the slopes in two temperature intervals, are shown in Table 4.2 and are compared with other published data.

Table 4.2. Comparison of activation energies and other performances of different SPADs

Ref	Tech. Node	Activation Energy E_a (eV)		SPAD Area (μm²)	Room temperature (~25°C)				Gating Rate max f_G, (MHz)	Gating Window t_{win}	Reset Time t_{RST} (ns)	Hold-off Time t_{off}	On-chip sync. and timing generation
		< −10°C	> 20°C		V_{BR} (V)	V_{ex} (V)	DCR (kHz)	DCR Density (MHz/mm²)					
[144]	350nm	0.15	0.57	400	28	4	4	10		10ns−25 ms	0.5−2	≤4μs	no
[145]	180nm	0.05	0.13	80	10.2	0.5	60	750		free-run	30	no	no
[146]	180nm			80 80 80	15 21 20.5	1.8 1.8 2.3	0.1 0.1 0.07	1.25 1.25 0.88	1 1 1	free-run	200	unused	no
[147]	130nm	0.09	0.15	80	9.7	1.7	100	1250				500μs	no
[148]	130nm		0.65	50	14.4	1.4	0.05	1	< 10	$1/f_G$-t_{RST}	> 20	unused	no
This work	**130nm**	**0.17**	**0.38**	**100**	**11.3**	**1.3**	**3**	**30**	**≤ 100**	**3.5 ns**	**0.5**	**$1/f_G$−4ns**	**yes**
[142], [143]	Si Photo-Multiplier	0.57 0.56 0.56	1.18 1.1 1.1	1600 1600 1600	28.3 ~30 ~30	4 3 3	6 4 0.56	3.75 2.5 0.35		free-run			no

First, observe in Table 4.2 that high DCR densities (MHz/mm^2) and low activation energies (E_a<E_g/2~0.56eV for silicon) are present in SPADs from DSM CMOS technologies, even at room temperature and above [145], [147]. The reduced activation energy indicates an increased density of non-mid-gap GR centers and a growing contribution from tunneling effect. Second, E_a is lower for SPADs with lower breakdown voltage (V_{BR}), since the tunneling effect is stronger in junctions with lower V_{BR} (high doping level and narrow depletion region). With the enhanced tunneling effect, note that the two high DCR densities are from SPADs with the lowest V_{BR} [145], [147]. Third, due to the reduced thermal generation, the activation energy at temperatures below -10°C is further reduced and tunneling becomes the dominant source of dark counts. Fourth, low DCR density and high E_a were achieved for SPADs from a devoted imaging fabrication process [148]. A Si photomultiplier is also provided at the bottom of Table 4.2. E_a of 1.1eV was achieved at higher temperatures, implying band-to-band generation and a high-quality fabrication process. However, these expensive processes are not easily integrated with the standard, low-cost digital and RF CMOS processes.

4.3.2.2 Afterpulsing Measurement

Afterpulsing refers to the secondary avalanche triggered by the released carrier which was trapped in deep energy levels during previous avalanche process. The time it takes to release a trapped carrier is related to the trap occupancy lifetime, which was found to be inversely proportional to temperature [149]. Generally, traps located at mid-bandgap energy levels have longer lifetime and therefore contribute more to afterpulsing. From the fabrication perspective, an efficient way to reduce the afterpulsing probability (AP) is to keep the manufacturing process clean, or use other special treatment of the chip to reduce the number of traps.

As introduced in section 4.1, AP can be effectively reduced by decreasing the number of filled traps, which is determined by the avalanche current density and its duration (quenching time). To reduce AP, high speed quenching circuits are usually designed to shorten the quenching time. The avalanche current, V_{ex}/R_D, can be reduced by using a lower excess bias V_{ex}, while the dynamic resistance (R_D) of the SPAD in breakdown is determined by the layers of the junction, thus R_D cannot be changed in a standard CMOS process. Also, V_{ex} cannot be reduced below large fraction of volt, since the sensing circuit in the SPAD pixel would become very complex. Therefore, active quenching is usually employed to reduce the quenching time, which reduces the avalanche charges and AP. For instance, fast active quenching and sensing circuits were designed in [150], [151], achieving low AP of 1.3% and 1.28% for hold-off times of 20ns and 5.4ns, but these fast quenching and reset circuits sacrifice the fill factor of the SPAD pixel.

The last option for reducing the avalanche charge is to restrict the available charge at the node of the cathode. For the proposed TG-SPAD, the avalanche charge is (V_{ex} x C_C), where C_C is the sum of all capacitances connected to the cathode node. Among the different contributions to C_C, the SPAD's capacitance is the largest, followed by the gate capacitance of the sensing circuit. The approach of restricting the available charge at the node of the cathode by means of minimizing the nodal capacitance is an efficient way to reduce AP in TG-SPAD. A detailed analysis of the nodal capacitance at the SPAD cathode in the TG-SPAD front-end circuit is presented below.

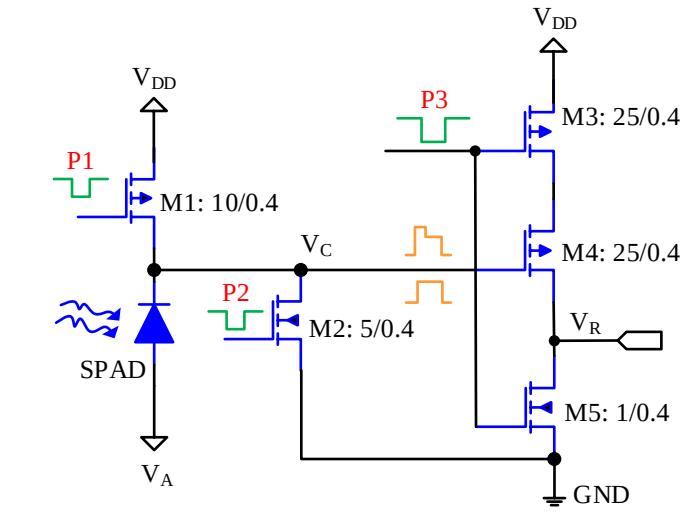
Figure 4.22(a) shows the schematic of the TG-SPAD front-end with the W/L ratios of all transistors indicated. The values in the W/L ratios are in micrometers. With the transistor models given in ref [152], the equivalent circuit of the TG-SPAD front-end is shown in Figure 4.22(b), from which the quenching time τ_{Quench} can be calculated from

$$\tau_{Quench} = R_D \cdot \left(C_{OXP1} + C_{OXN2} + C_D + \frac{3}{2} C_{OXP4} \right). \tag{4.11}$$

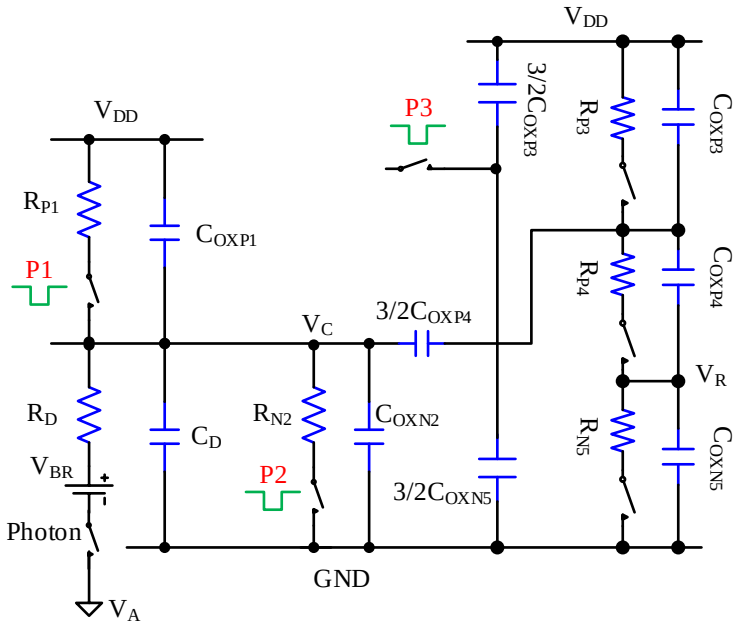
The C_{OX} terms in Eq. (4.11) are the gate oxide capacitance of the transistors, which can be written as

$$C_{OX} = C_{OX}^{'} \cdot WL\,. \tag{4.12}$$

$C_{OX}^{'}$: Gate oxide capacitance per area



(a)



(b)

Figure 4.22: Schematic (a) and equivalent circuit (b) of the TG-SPAD pixel circuit

The gate oxide thicknesses of NFET and PFET in this process are 59Å and 61.5Å respectively. Therefore, the gate oxide capacitances per area of NFET and PFET are 6fF/μm^2 and 5.7fF/μm^2, respectively. With the transistor W/L ratios given in Figure 4.22(a), total parasitic capacitance of the SPAD is 220fF, with contributions of ~35fF from M1 and M2, ~85fF from M4, and ~100fF from the diode itself. In addition, the breakdown resistance R_D of the SPAD was determined to be 380Ω by Eq. (4.7) in section 4.2.1, so the quenching time of this TG-SPAD is short (< 90ps), even though an active quenching circuit is not used.

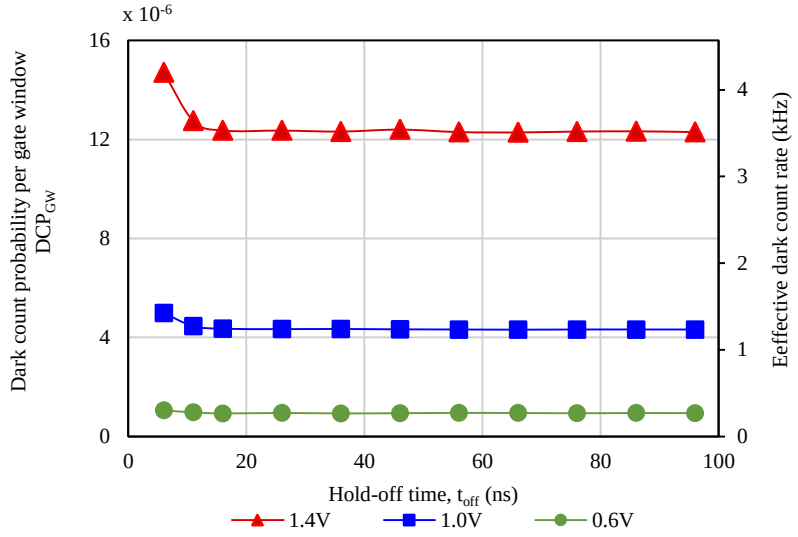


Figure 4.23: Measured dark count probability per gate with deadtime

As explained above, AP is caused by the emission of trapped charge, captured from preceding avalanche of charge multiplication. The release of a trapped charge is determined by its lifetime. An advantage of TG-SPAD is that by changing the frequency of the external triggering signal, different hold-off times can be applied to wait the trap emission to be completed, and then activating the gate window. In this way, afterpulsing can be controlled. It is possible to use the dark count measurement to evaluate AP. The assumption is that dark counts originate from a time-invariant generation of carriers in the SPAD, either due to thermal generation or from tunneling. By varying the hold-off time of the TG-SPAD, it is expected that the dark count rate stays unchanged, but the afterpulsing probability changes. Thus, to evaluate the AP of this design, DCP_{GW} was measured as functions of the hold-of time t_{off} and excess bias V_{ex}, as shown in Figure 4.23. The effective DCR is reflected in the right-hand axis of Figure 4.23.

For a fixed excess bias, DCP_{GW} is virtually the same for $t_{off} \geq 16$ns. However, due to afterpulsing, DCP_{GW} increases noticeably when $t_{off} \leq 11$ns. The corner $t_{off} \sim 11$ns indicates that afterpulsing in this TG-SPAD is extinguished, when holding off the SPAD below breakdown with transistor M2 for 11-16ns. It is a very useful feature of the time gating function in this front-end. As given in [153], AP can be calculated by Eq. (4.13).

$$AP = \frac{DCR\text{-}DCR_0}{DCR}. \tag{4.13}$$

In Eq. (4.13), DCR_0 is the DCR without afterpulsing, measured at long hold-off times, $t_{off} > 50$ns. Using Eq. (4.13), AP is calculated as function of t_{off} for V_{ex} =1V and the results are given in Figure

4.24. Here, it is seen that AP is low and negligible for longer hold-off times (t_{off}≥16ns). Since the DCR does not change for hold-off times longer than 50ns, AP equals to zero for these longer hold-off times. For comparison, values for AP from other technologies or pixel designs are also shown in Figure 4.24.

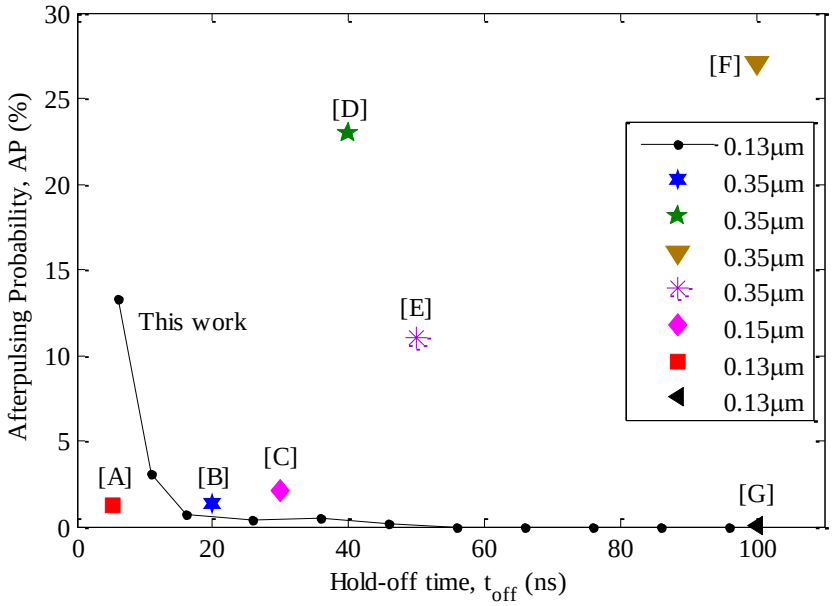


Figure 4.24: Measured afterpulsing probability of the TG-SPAD (line with small circles), and comparison with reported data (symbols only) from other publications (Key to references: A [150], B [151], C [154], D [155], E [156], F [157], and G [148])

4.3.2.3 Photon Detection Efficiency (PDE) Measurement

The photon detection efficiency (PDE) measures the probability of an incident photon to be detected by the SPAD. In order to be detected, a photon should first reach the depletion region layer of the SPAD, then the photon energy should be absorbed in the semiconductor to generate electron-hole pair of primary carriers. Finally the primary carriers have to successfully trigger an avalanche. These processes are all with probabilities less than unity. For example, the photon must pass through the entire stack of passivation and dielectric layers, and the photon can be reflected by the material interfaces in the stack or absorbed. Thus, PDE is the product of the transmission coefficient of the passivation layers, photon absorption efficiency of SPAD, and the probability of a primary carrier to trigger an avalanche multiplication.

A block diagram of PDE measurement is shown in Figure 4.25. The incident light source was a pulsed laser (PicoQuant: LDH-D-C-510, 130ps pulse width, 510nm, 80MHz), driven by a laser driver electronics (PicoQuant: PDL-800-B). To synchronize the laser pulse with the 3.5ns gate

window, a delay unit (Optronics-TRRC1) was used to adjust the time position of the pulsed laser to ensure that the laser pulse is within the gate window.

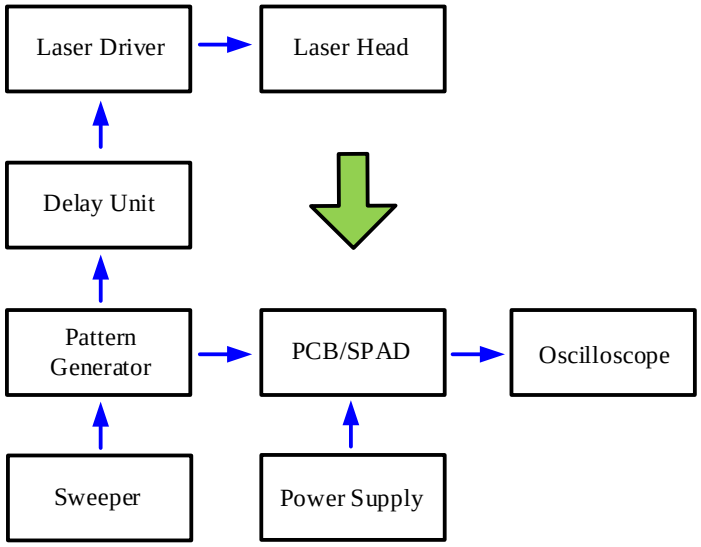


Figure 4.25: Schematic diagram of the PDE measurement

PDE of the TG-SPAD front-end is calculated with Eq. (4.14) [158], where μ is the number of incident photons per laser pulse, f_{GW} ($1/t_{win}$) is the frequency of the gate window, C_{Dark} is the number of counts in dark, and $C_{Illumination}$ is the number of counts with illumination.

$$PDE = \frac{1}{\mu} ln \frac{1 - C_{\text{Dark}} / f_{GW}}{1 - C_{\text{Illumination}} / f_G} \tag{4.14}$$

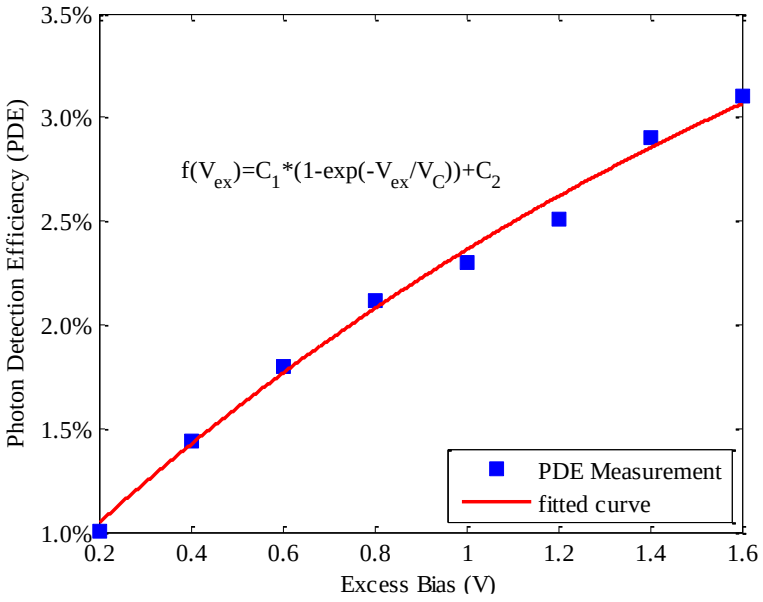


Figure 4.26: Measurement of PDE as a function of excess bias

Figure 4.26 shows the measured PDE as a function of V_{ex}, from which we can see that PDE increases with V_{ex}. This is because that the avalanche triggering probability (ζ) is strongly bias dependent ($\zeta \propto (1-e^{-V_{ex}/V_C})$) [149], [159]. V_C is the characteristic voltage, having a value of a few volts for SPADs with thin depletion regions [159]. The curve fitting result of the PDE measurement is also shown in Figure 4.26, with V_C equals to ~2.3V. Comparing with the SPADs implemented in high-voltage processes, or the optimized imaging process, PDE of this TG-SPAD is lower. However, this is expected for a SPAD implemented in DSM CMOS technology, owing to the narrow depletion region and strong surface reflections.

4.4 Summary and Future Improvements

The TG-SPAD front-end has been characterized from performances measurements of dark count probability per gate window (DCP_{GW}), afterpulsing probability, and PDE. It was shown that an important benefit of time-gated operation is the reduction of both dark counts and afterpulsing probabilities. However, because of the dielectric layer stack and design rules of this inexpensive CMOS technology, both fill factor and PDE of the TG-SPAD front-end are relatively low.

An advantage of the TG-SPAD front-end is that it contains only 5 transistors. The simple circuitry of this design is beneficial for improving the pixel fill factor. The main reasons for the low fill factor are in the design rules for the layout of the structure. As shown in Figure 4.7, for example, there is minimum spacing of the deep N-well region (2.2µm). That is, for a 10µm x 10µm square-shaped diode, there is at least 4.4µm spacing around the diode, causing the fill factor of the diode alone to be only 16%. The fill factor of the SPAD can be improved by increasing the pixel size, which is not favorable of array designs for miniaturized spectrometers. Also, with larger areas SPADs, the dark count probability increases.

Concerning PDE improvement, the main challenge is from the surface passivation layers, which are used to protect the device. Since removing of passivation layers is not an option in standard CMOS, some other methods for post processing of the chip surface can be tried. Possible methods are deposition of antireflection coating, selective etching of chip areas, or attachment of a micro lens array. However, post-processing will increase the cost of the SPAD.

Chapter 5

Time-Gated Spectrometer Implementation and Applications

In this chapter, the prototype of the time-gated spectrometer is discussed. The time-gated spectrometer combining the wavelength selector and TG-SPAD is built. The setup and synchronization of the time-gated spectrometer is described in section 5.1, followed by a system characterization, and examples of two major applications for Raman spectra and fluorescence lifetime measurements.

5.1 Time-Gated Spectrometer Implementation

5.1.1 System Configuration

To build a time-gated spectrometer, selection of main system components and data acquisition method must be carefully considered. As described in chapter 2, the excitation source plays an important role in spectrum quality, because it determines the stability and intensity of the Raman signal. Taking into account the low Raman scattering efficiency, the low-power pulsed laser used in section 4.3.2.3 is not suitable for excitation of Raman signal. Instead, a high-power solid state pulsed laser (Passat Compiler 355) is used.

The specifications of the selected pulsed laser are given in Table 5.1. This laser is multifunctional and equipped with three output channels for emission at three wavelengths, 355nm, 532nm, and 1064nm. The pulse width of the 532nm channel is 7ps, and the maximum repetition rate is 200Hz. This repetition rate limits the maximum achievable counting rate of the detector when it is used in time-gated applications. In order to simplify the setup, light is transmitted and collected by a multimode optical fiber (Ø200μm, 0.22 NA).

Table 5.1: Specifications of Passat Compiler 355 pulsed laser

Parameter	Vaule
Wavelength	355 nm, 532 nm, 1064 nm
Repetition Rate	Internal/External triggering, 200 Hz maximum - variable from 1 Hz to the maximum via RS-232 port
Energy Output (at 100 Hz)	150 μJ/pulse at 355 nm 160 μJ/pulse at 532 nm and 1064nm
Pulse width (at 532 nm)	7ps
Beam profile	Close to Gaussian
Beam Diameter	~ 1.2 mm
Line width	~ 2.7 cm^{-1}
External control	Connector for TTL trigger input or +4 +/-1V, into 1kΩ

In principle, the function of a spectrometer is to measure the spectrum of an input signal. This spectrum can be measured either by a monochromator or by a multichannel spectrometer. In a monochromator, a discrete detector is used for light detection, and the spectrum covering the entire wavelength band is obtained by rotating the grating to scan the wavelength band. In a multichannel spectrometer, a detector array with dimension of several millimetres is used to simultaneously detect the spectrum of all wavelength components.

The Raman shift for most chemicals is in the range of 500-3000cm^{-1}, which implies that the Raman spectra wavelengths are few tens to hundred nanometers around the laser excitation wavelength of 532nm. For Stokes shift, the wavelength band of the Raman spectrum is from 546nm to 633nm (Table 2.6). Also, the Raman signal is delivered to the spectrometer by the fiber mentioned above. From chapter 4, a single pixel TG-SPAD front-end was designed, with square-shaped active area of 10μm x 10μm. Thus, the system is suitable for operation as a monochromator, with input slit of 0.2mm and output slit of 0.01mm. The concave grating designed in chapter 3 can be used to perform the wavelength separation function.

In a planar grating based monochromator, selection of the output wavelength is achieved by adjusting the incident angle. However, in a concave grating based system, owing to the focusing property of the Rowland configuration, both the input and output slits must be on the virtual Rowland circle of the concave grating. Thus, a simple rotation of the concave grating is not sufficient for wavelength selection, since the Rowland circle rotates around a point on the circle, rather than around the circle center. Therefore, a rotation of concave grating is not desirable, since by rotating the concave grating, both the input and output slits will move away from the Rowland circle.

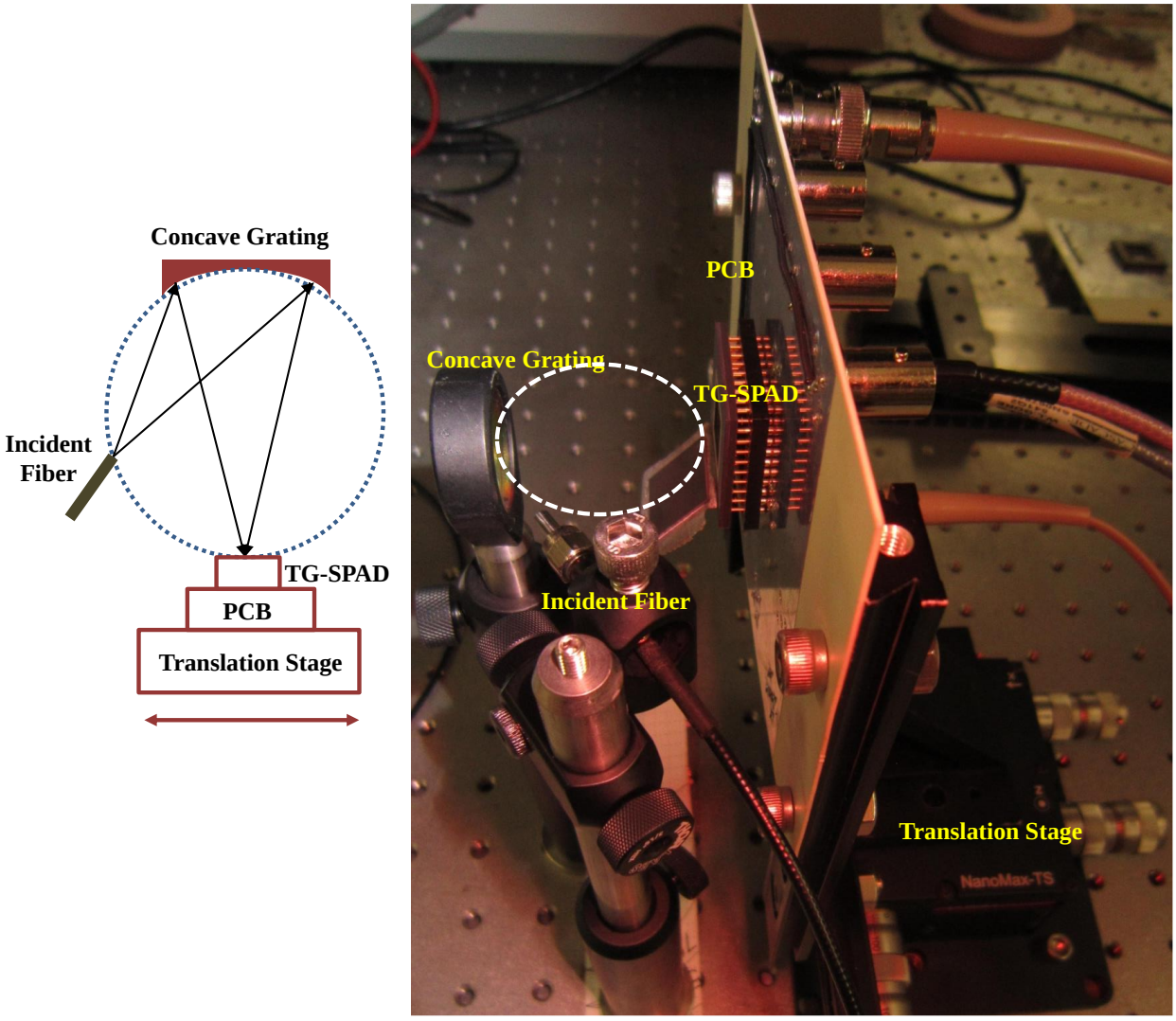


Figure 5.1: Optical setup of the system in a Rowland configuration

To build a monochromator with a concave grating, the incident angle can only be adjusted by moving the incident slit along the Rowland circle. However, the input slit is a fiber, and moving a fiber can introduce instability or perturbation to the system. For repeatability of the spectrum acquisition, the incident fiber must be fixed, and instead the detector was moved to acquire data at different wavelength positions. The TG-SPAD was mounted on a 3-axis translation stage. Figure 5.1 shows the experimental setup of the system in Rowland configuration. The PCB with the TG-SPAD was mechanically reinforced by metal bars and the bars were bolted firmly on the translation stage (Thorlabs, NanoMax TS 313D), which offered 4mm travel range in each axis with coarse and fine adjustments of resolutions of 10μm and 1μm, respectively.

5.1.2 System Synchronization

The synchronization between the pulse excitation and the acquisition window is critical for the operation of a time-gated system. This is to ensure that the optical signal from the target and the narrow gate window of the detector coincide in time. Otherwise, the optical signal can easily be missed in the time-gated measurement.

The time-gated spectrometer is designed to measure Raman spectra emitted simultaneously with the optical excitation. Thus, it is necessary that the 7ps laser pulse and the 3.5ns gate window of the detector be synchronized. The setup shown in Figure 4.25 was tried by replacing the low-power laser with the high-power solid-state laser. Unfortunately, no photon was detected even if large delays were used. Perhaps this is caused by the jitter of the laser triggering, because the maximum repetition rate of the laser is only 200Hz. At this low repetition rate, possibly the jitter between laser trigger signal and laser emission is considerably larger than the 3.5ns gate window. Therefore, it might be impossible to find a single delay to fix the narrow laser within the 3.5ns gate window. To verify this hypothesis, the time jitter of the pulsed laser triggering was tested.

Figure 5.2(a) shows the block diagram of the laser triggering jitter measurement. A clock signal was provided from a pulse generator (Quantum 9520 Series) and used to simultaneously trigger the pulsed laser and the oscilloscope (Lecroy). The emitted laser pulse was detected with a high speed photodiode (Thorlabs, DET10A). The photodiode produced a step voltage with a rising edge of 1ns at each laser pulse. The build-in functions in the oscilloscope were used to measure the time interval between the triggering signal and the rising edge of the signal from the photodiode. A histogram of the measured time interval is plotted in Figure 5.2(b). From the histogram range (the x-axis), the triggering of the laser takes about 158μs, but the triggering time is broadly distributed within 500ns (measured standard deviation ~160ns), which indicates that the triggering jitter of the laser was considerably larger than the acquisition window of 3.5ns. Thus, the simple global synchronization in Figure 4.25 is abandoned, due to jitter in the laser triggering, and instead, a local synchronization of the photodetection to the optical excitation was pursued.

The local synchronization of the photodetection to the optical excitation can be termed simply as optical synchronization of the time-gated spectrometer. The triggering and synchronization are illustrated in Figure 5.3. The pulsed laser was externally triggered at 200Hz repetition rate. The photodetection was synchronized to the laser pulse, rather than to the pulse generator, as follows. The emitted laser pulse was split into two beams (30:70) by a beam splitter. The 70% channel was

used as the excitation source in the Illumination Channel of the time-gated spectrometer, while the 30% channel was detected by a high speed photodiode in the Synchronization Channel. Upon occurrence of a laser pulse, the photodiode generates a voltage step, which triggers the TG-SPAD. The photodiode was placed close to the beam splitter and the cable from the detector to the TG-SPAD was of relatively short length, so the delay in the Synchronization Channel was small, ~1ns.

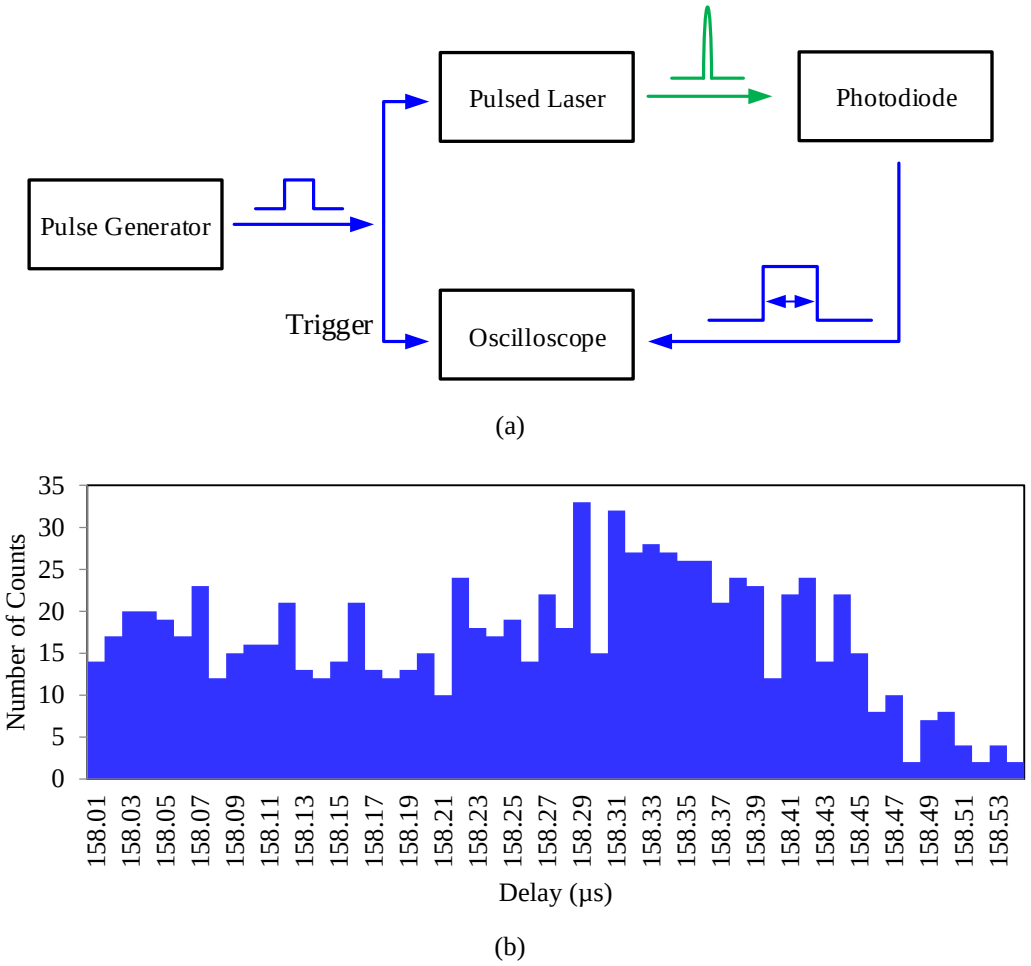


Figure 5.2: (a) Schematic diagram of the time jitter measurement; (b) Histogram of the time between laser trigger signal and laser emission

The delay in the Illumination Channel was higher, since the multimode fiber was several meters long. To equalize the delays in the optical and electrical paths, a delay unit was inserted between the photodiode and the TG-SPAD. Changing the delay of the delay unit, the gate window of TG-SPAD can be adjusted to precede or lag the arrival of the light pulse from the Illumination Channel. Positive pulses can be generated at the output of the TG-SPAD when photons arrive within the gate window.

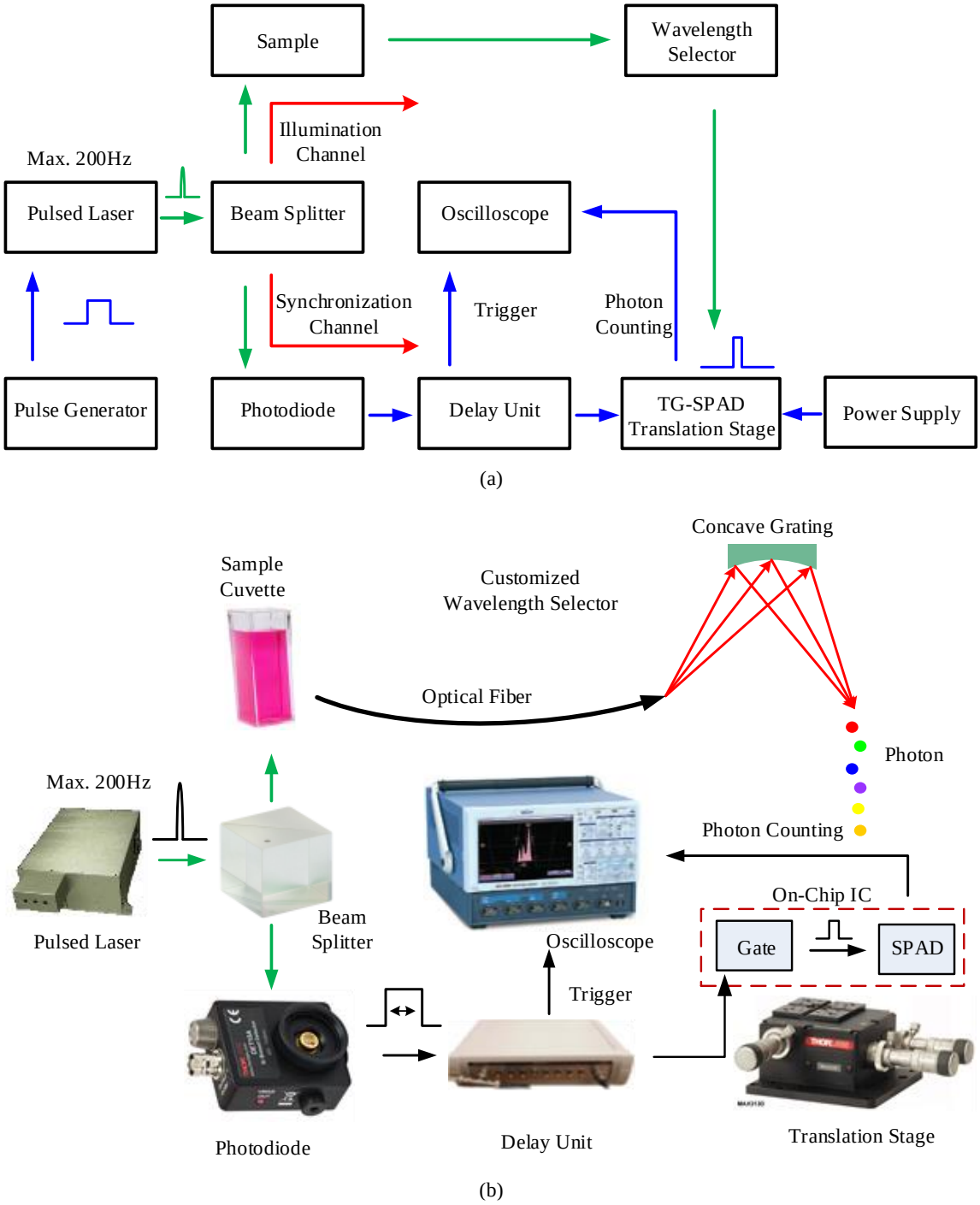


Figure 5.3: (a) Block diagram; (b) experimental setup of the optically synchronized time-gated spectrometer

The high-speed oscilloscope acquires the pulse from the TG-SPAD, recognises the pulse occurrence (by amplitude threshold) and calculates statistics. The statistical calculation includes number of pulses and histogram of the pulse width. Upon completing the measurement, the

oscilloscope saves the statistical data on disk. Then the data were transferred to a computer for processing and documenting.

It should be noted that the optical synchronization causes a minimal overhead of one beam splitter and one fast photodiode, when compared with the global synchronization. However, the overhead is cost effective, taking into account the large and very expensive overhead for making a jitter-free triggering of picosecond pulsed lasers. In addition, the optical synchronization allows for easily changing the pulse laser, without reworks and adjustments that the global synchronization requires.

5.1.3 System Timing Resolution

An important parameter of a time resolving system is the timing resolution, because it reflects the system's timing accuracy. The setup shown in Figure 5.3 has overcome the jitter in the pulsed laser triggering, but several other sources of time jitter remain. These sources of jitter are from the high-speed photodiode (11ps, [160]), delay unit, oscilloscope (<2.5ps), and the TG-SPAD. The square root of quadratic sum of the jitter contributions from each source yields the overall system timing resolution.

$$FWHM = \sqrt{\sum_{i=1}^{n} FWHM_i^2} \tag{5.1}$$

Figure 5.4 shows the histogram (blue color) of the measured photon arrival time (blue) of the 7ps pulsed laser within the gate window. Approximating the histogram with a Gaussian fitting (red curve), indicates that FWHM of the photon arrival time is 60ps. Taking into account the contributions from all jitter sources, the timing resolution of this TG-SPAD is better than 60ps.

Timing jitter refers to the temporal correspondence between the arrival of a photon and the detection of a resulting avalanche. A significant contribution of timing jitter is the drift (or diffuse) of photo-generated carriers from the absorption point to the high field depletion region and subsequent triggering of an avalanche. There is statistical fluctuation of delay between photon absorption and avalanche triggering, contributing to the SPAD timing jitter. Timing jitter of a SPAD was proposed to be dependent on position statistics of photon absorption [161]. It was deduced in [162], that in principle, an SPAD with smaller active area and narrower depletion width should have better timing resolution.

A simple comparison in Table 5.2 confirms the relation between SPADs from different technologies. Comparing the performance of SPADs in Table 5.2, the TG-SPADs in this work possess better timing resolution and lower power consumption, owing to their small areas and narrow depletion regions. However, the narrow depletion region and the polyimide layers on top of the active region in the standard 130nm CMOS process reduce the PDE, compared with PDE of SPADs fabricated in high-voltage CMOS with wider depletion layers. Thus, a tradeoff between timing resolution and photodetection efficiency is evident.

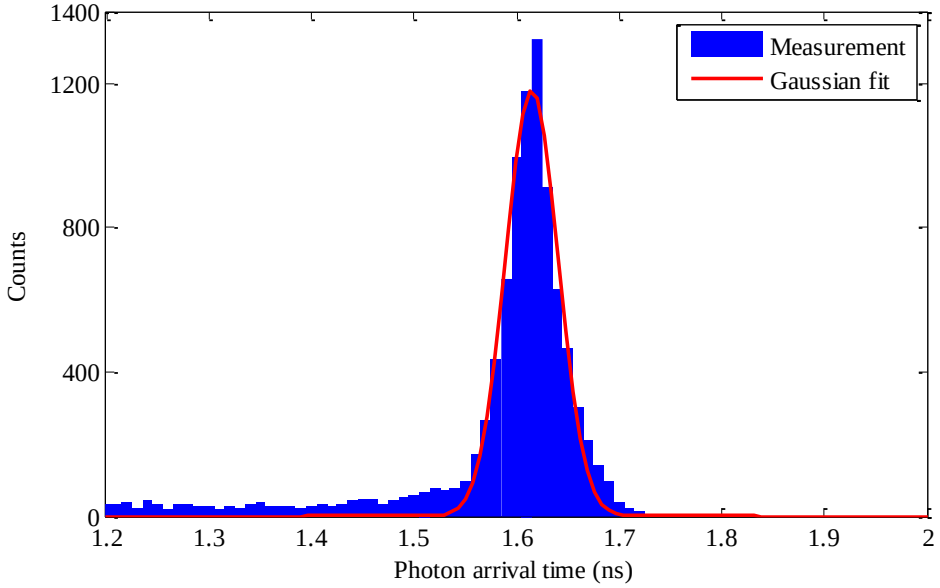


Figure 5.4: Photon arrival time measurement of a 7ps pulsed laser

Table 5.2: Comparison of performance characteristics and applications of state-of-the-art CMOS TG-SPADs

	This work	[127], [128]	[126]	[156]
Technology	CMOS 0.13µm	CMOS 0.35µm HV	CMOS 0.35µm HV	CMOS 0.35µm HV
Number of pixel	1	1	1024 x 8	10 x 43
Detector area	10µm x 10µm	10µm x 10µm 20µm x 20µm	24µm x 24µm	20µm x 100µm
Fill factor	9.8%	4% or 11%	44.3%	67%
PDE	3%@510nm (V_{ex}=1.4V)	30%@500nm	9.6%@465nm (V_{ex}=3V)	4%@500-700nm (V_{ex}=1.0V)
DCP_{GW}	5×10^{-6} (V_{ex}=1.0V)	3.75×10^{-6} (V_{ex}=2.3V)	4×10^{-6} (V_{ex}=3V)	2×10^{-4} (V_{ex}=1V)
Gate window	3.5ns	0.45ns	0.7ns	4ns
Timing resolution	60ps	100ps	250ps	
Application	Raman	Raman	Raman	2D Imager

5.2 Raman Spectrometer Verification

5.2.1 Raman Spectrum Measurements

When measuring the low-intensity Raman spectrum, the requirement for a minimum intensity of the excitation source is based on the detection limit of the system, especially that of the detector. From Table 5.2, the DCP_{GW} of the detector is 5 x 10^{-6}, which implies the probability to have Raman photon detected per gate should be larger than 5 x 10^{-6}. Considering a signal-to-noise ratio (*SNR*) of 40, then the minimum Raman photon detection probability per gate P_{min}=5 x 10^{-6} x *SNR* =2 x 10^{-4} = 0.02%. A more practical interpretation is that at least one Raman photon should arrive at the TG-SPAD in $1/P_{min}$=5000 gate windows in order to achieve a *SNR* of 40.

Since P_{min} is a small fraction of unity, then one can be easily misled that there is no problem to have the small number of Raman photons after excitation with a pulsed laser. However, it is necessary to consider all optical attenuations of the laser excitation in order to prove that at least P_{min} Raman photons actually arrive at the TG-SPAD in the gate window.

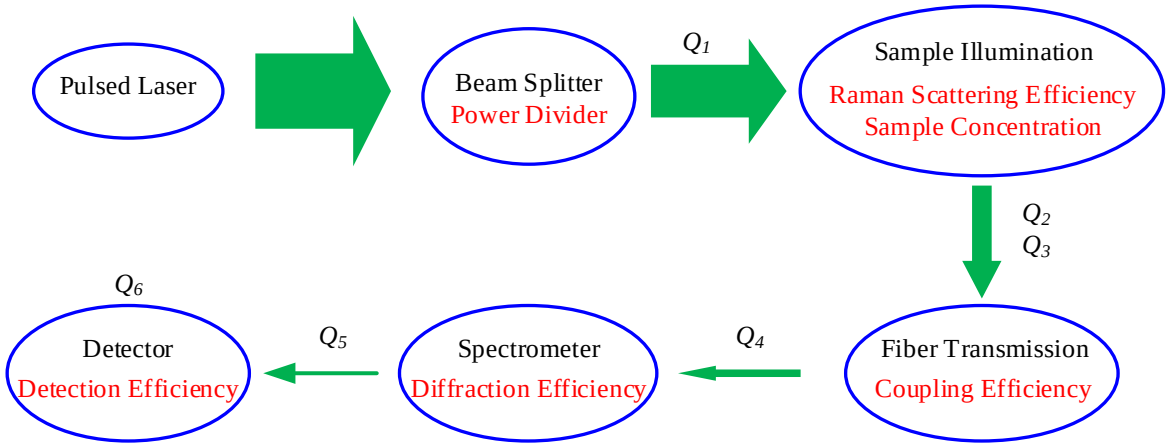


Figure 5.5: Light attenuation in Raman spectrometer

Figure 5.5 details the light attenuation in the time-gated system given in Figure 5.3. The incident light is split by a beam splitter fist, and only 70% (Q_1) of the laser light is transmitted for sample illumination. The second loss comes from the Raman scattering process. It was reported that only one Raman photon can be generated for every 10^{10} incident photons [36], which indicates a very weak scattering process with efficiency Q_2=10^{-10}. Since the sample is dissolved in water, so sample concentration also reduces the Raman photon rate (Q_3=1%). The scattered light is collected by a multimode optical fiber, and the coupling efficiency depends on the fiber's numerical aperture. In addition, light is scattered in all directions, and the sample-to-fiber coupling (Q_4=0.1) is further

reduced. In the wavelength selector, additional light loss is present for the 1st order diffraction of the concave grating (Q_6=0.2). The Raman signal is finally arriving at the TG-SPAD, with PDE of Q_6=3%. To achieve the predetermined *SNR*, eventually the detection probability of Raman photon per gate should larger than P_{min}.

With all of the losses considered, and to achieve Raman photon detection with SNR≥40 by the detection probability per gate P_{min}=0.02%, the number N_0 of photons in each laser pulse should meet the following condition

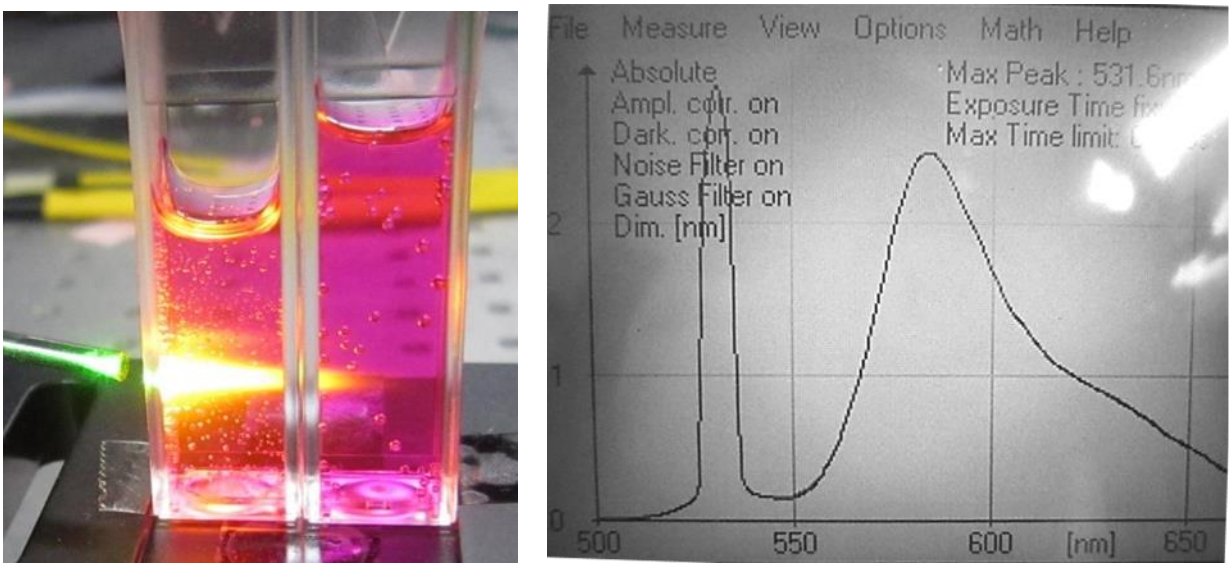
$$N_0 \cdot Q_1 \cdot Q_2 \cdot Q_3 \cdot Q_4 \cdot Q_5 \cdot Q_6 \geq SNR \cdot \mathrm{DCP_{GW}} = 2 \times 10^{-4} \tag{5.2}$$

Substituting all the losses into Eq. (5.2), the minimum incident photon number is calculated to be 5 x 10^{12}. Since the energy of a photon at 532nm is ~3.7 x 10^{-19}J, so the minimum energy of the laser pulse equals to ~20μJ/pulse. The selected high-power pulsed laser met the energy requirement, as seen in Table 5.1, which allowed for the successful Raman measurement of an organic sample. Thus, although the detection limit of the TG-SPAD is very good, P_{min}~1photon per 5000 gate windows, it still needs to have high-energy laser excitation for measuring a Raman spectrum. This is because of the various light losses in the setup, while the most significant loss being the low efficiency of Raman scattering.

In addition to the energy required from the excitation source, the other issue in the Raman measurement is fluorescence of the sample. The most significant feature of the proposed system is the very high probability to suppress the fluorescence signals. Therefore, to test the efficiency of this time-gated system, a dye of Rhodamine B with strong fluorescence emission is used. This dye was selected according to the wavelength of the laser (532nm), and its excitation (562nm) and emission peaks (583nm) [163]. The purpose of this experiment was to test if the system can measure the Raman spectrum when a strong fluorescence signal was present. For comparison, the spectrum was also measured by a general purpose spectrometer (OEM-400, Newport), operating in free running mode.

During the measurements, Rhodamine B was dissolved in water and carried in a plastic cuvette. As shown in Figure 5.6(a), the sample cuvette was illuminated by the pulsed green laser, and the emitted fluorescence signal of yellow color was observed. Two cuvettes were used, the one on the right was used to block the transmitted excitation signal. The emitted signal was collected by a fiber and measured with the commercial spectrometer. Figure 5.6(b) shows the measured spectrum of Rhodamine B, from which two strong peaks are observed. The first peak is the Rayleigh

scattering at the wavelength 532nm of the excitation source. The second broad peak is at ~580nm and matches the fluorescence emission band. Notice that no Raman peak was resolved by the free running commercial spectrometer, because the weak Raman signal is easily overwhelmed when a strong fluorescence is present.



(a) (b)

Figure 5.6: (a) Illumination of sample carried in plastic cuvette; (b) Spectrum measured by a commercial spectrometer (OEM-400, Newport, Irvine, CA, 0.3nm spectral resolution)

In a time-gated measurement system, the width of the gate window is important for suppression of fluorescence. From chapter 4, the 3.5ns gate window of the TG-SPAD is fixed by on-chip pulse generators. The fluorescence lifetime of Rhodamine B was reported to be ~1.7ns [164] when dissolved in water, which implies that the majority of the fluorescence signal would be emitted within a 3.5ns gate window. In addition, the emission peak of the fluorescence at 583nm overlaps with the major Raman peaks of Rhodamine B. Thus, the selection of a short width <1.7ns of the gate window should effectively suppress the fluorescence in a time-gated measurement.

To shrink the 3.5ns gate window of the TG-SPAD into a shorter *detection window*, the TG-SPAD was triggered before the arrival time of optical signals from the target. As explained in Figure 5.3, the delay in the Illumination Channel is longer than the delay in the Synchronization Channel, and the triggering of the TG-SPAD before the arrival of the optical signal at the SPAD is accomplished by reducing the delay in the delay unit in the Synchronization Channel. As shown in Figure 5.7, the *detection window* was defined as the time between the arrival of the excitation signal at the SPAD and the end of the gate window. Since both the Raman and fluorescence emissions

are stimulated by the excitation laser pulse, so no photon is present before the *detection window*. Therefore, the *detection window* can be made shorter than the 3.5ns gate window. To adjust the width of the *detection window*, a delay unit was used to change the delay between the start of the 3.5ns gate window and the excitation. The dependence of fluorescence suppression on the width of the *detection window* was investigated.

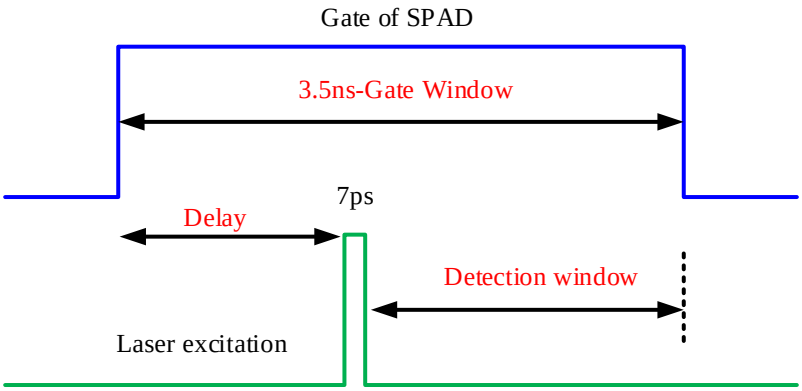


Figure 5.7: Timing diagram of the measurements

Having in place the setup of the time-gated spectrometer with a technique for fluorescence suppression by means of short *detection window*, the emission spectra of Rhodamine B were measured at several different and short *detection windows*, as shown in Figure 5.8. In more detail, the sample was under 532nm excitation, and emission from the sample was delivered on the Rowland circle of the concave grating. The chip with the TG-SPAD was also aligned on the Rowland circle (@532nm Rayleigh scattering position) of the concave grating, and then moved in steps of 10−50μm on the tangent of the circle to acquire the data for the spectrum range from 520nm to 600nm. This wavelength range corresponds to Raman shifts up to 2100 cm^{-1}.The spectrum of Rhodamine B was measured under 6 different *detection windows*, of widths from 3ns down to 250ps. The data for each position were collected from 10 000 laser shots. The dark count in these 10000 acquisitions was less than 1, on average, and it was negligible compared with the number of counts in the spectra. The spectra, plotted as number of counts vs. the Raman shift, are shown in Figure 5.8.

5.2.2 Analysis of the Time-Gated Raman Spectrum Measurement and Discussions

Similar to the free running spectrometer measurement shown in Figure 5.6(b), a strong Rayleigh scattering peak was resolved. The intensity of the Rayleigh scattering did not change much with the detection window, because of its instantaneous response to the excitation.

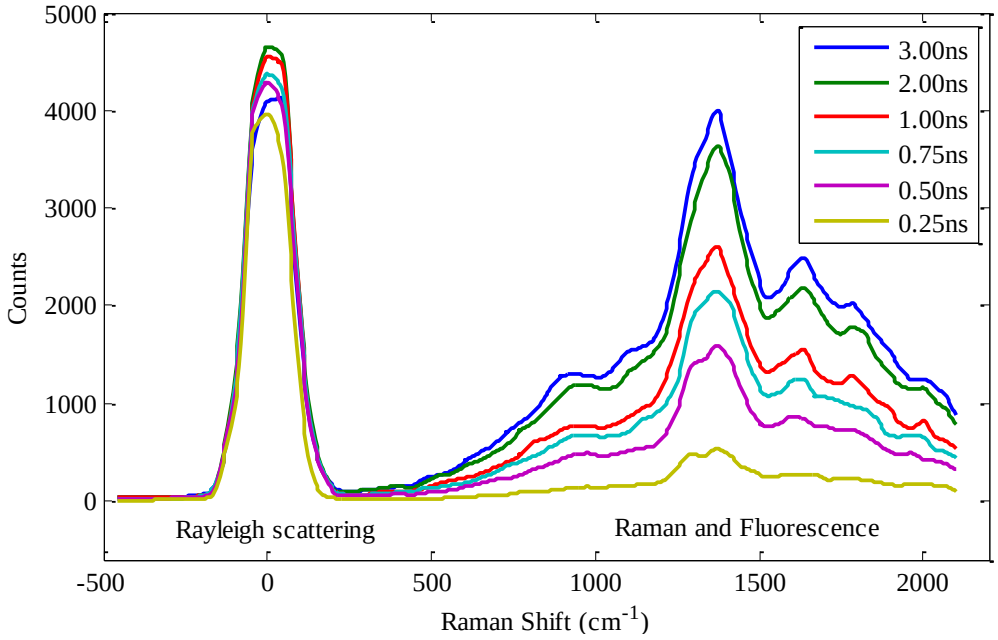


Figure 5.8: Spectrum of Rhodamine B measured with different detection window

A strong fluorescence emission in a broad band was also observed for the *detection window* of 3ns. However, in contrast to Rayleigh scattering, the intensity of the fluorescence emission is significantly decreased by narrowing the *detection window*, proving the efficiency of narrowing the *detection window* to suppress fluorescence signals.

In addition to the broad fluorescence emission band, and in contrast to the free running spectrometer measurement, the time-gated measurement has acquired a strong Raman peak, superimposed on the broad spectrum of the fluorescence signal, as seen in Figure 5.8 for the 3ns *detection window*. Actually, this peak (@~1300cm^{-1}) is a superposition of two peaks, since the peak at *detection window* 3ns gradually splits into two peaks (dashed block in Figure 5.8) when the *detection window* is reduced to 250ps. Figure 5.9(b) gives a better perspective for the evolution of the Raman peaks as function of the *detection window*.

To obtain the positions of the Raman peaks, the measured data from Figure 5.8 were processed as follows. First, the fluorescence background was extracted, and the background levels (curves in Figure 5.9(a)) decreased as the *detection window* was reduced. Second, the background was subtracted[2] to obtain the Raman spectra (symbols) in Figure 5.9(b). Third, the Raman spectra were fitted with Gaussian fitting, curves in Figure 5.9(b), and from the fitting results, the Raman shift and FWHM of the Raman peaks were obtained. The vertical dashed lines in Figure 5.9(b) depict

[2] Background correction—Backcor, MATLAB central

the average positions (6 *detection windows*) of two Raman peaks, R1 and R2. Stable values for the Raman shift of these peaks R1 and R2 were observed at different *detection windows*, since the two Raman peaks were well aligned with the vertical dashed lines. The average values of R1 and R2 are 1372cm^{-1} and 1284cm^{-1}, respectively, which are in good agreement with the results (1365cm^{-1}, 1290cm^{-1}) measured for Rhodamine B by a commercial surface enhanced Raman spectrometer [165]. The good agreement (within ± (6−8) cm^{-1} between this and published data) verifies that the proposed spectrometer with concave grating and TG-SPAD is suitable for Raman spectroscopy.

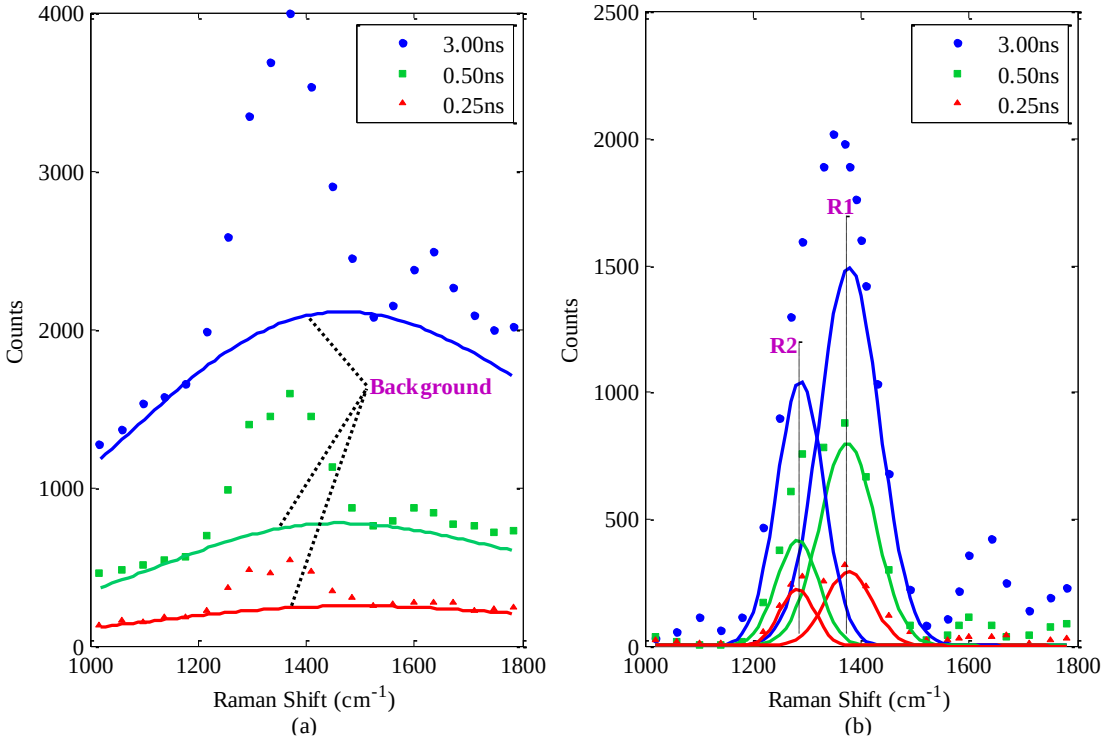


Figure 5.9: Fluorescence background (a) and Raman peaks (b) measured at different detection windows

Another observation in Figure 5.9(b) is the FWHM of the Raman peaks, in the order of 100 cm^{-1} for longer detection windows, as deduced from the fitted curves in Figure 5.9(b). As discussed in chapter 3, the entrance slit (width and numerical aperture) plays an important role in system's spectral resolution. To achieve high spectral resolution, narrow entrance slits (10µm-25µm) are employed in commercial spectrometers. However, the entrance slit width of this setup is determined by the multimode fiber diameter (200µm). The large diameter fiber was chosen through a trade-off between spectral resolution and light intensity that can be coupled and transmitted, since high intensity is also important for the measurement. In addition to the wide entrance slit, the curved focal plane of the grating also contributes to the low spectral resolution. Because the SPAD

was linearly moved on the tangent of the Rowland circle by a translation stage, then some fraction of the spectrum was measured at positions out of the Rowland circle, which degrades the spectral resolution.

In addition, the limit of detection (LOD) of the system is important. The concentration of the sample measured above was 10*m*mol/L. According to [166], the intensity of a Raman peak is directly proportional to the analyte concentration. Therefore, if the sample is diluted 10 times, then the peak intensity in Fig. 5.9(b) will also be attenuated by a factor of 10. The LOD of the system refers to the lowest sample concentration at which the Raman signal is distinguishable from the background signal. Assuming the lowest signal intensity that can be distinguished is 10 times the background signal, then for the *detection window* of 250ps, the lowest sample concentration is (10*m*mol/L / (279/10/background)) 17μmol/L, where 279 is the signal intensity at 250ps *detection window* (Fig. 5.9(b)).

Compared with the high resolution but also expensive commercial Raman spectrometers, the custom prototype of this miniaturized Raman spectrometer failed to detect the weak Raman peaks, owing to poor spectral resolution. However, the low-cost concave grating together with the TG-SPAD fabricated in mainstream CMOS, are beneficial to decrease the cost and size of the Raman spectrometer, providing also a fluorescence suppressed spectrum, and these features are of great importance for field applications of Raman spectroscopy.

5.3 Fluorescence Lifetime Measurements

5.3.1 Fluorescence Decay Measurements

In addition to Raman spectroscopy, another application of the time-gated system is fluorescence lifetime measurement. The experimental setup of the fluorescence lifetime measurement is the same as that shown in Figure 5.3.

The principle of the time-gated fluorescence lifetime measurement is shown in Figure 5.10 and the inset figure depicts the temporal distributions of excitation and fluorescence emission. The fluorescence signal is emitted in a large time scale and decays exponentially. This time constant of the exponential decay is often termed the fluorescence lifetime.

As shown in Figure 5.10, to measure the fluorescence emission decay, the gate window of the TG-SPAD is placed at different delays with respect to the excitation. A delay generator was used to adjust the delay between the excitation and the gate window of the detector, with steps between

250ps and 2ns, depending on the change of signal intensity. Also, well distinguishable differences in the counts from the TG-SPAD front-end were present at selected delays. The time-resolving histograms of counts for each setting of the delay line were collected from 10 000 laser shots. The fluorescence lifetime was then extracted from the decay curve.

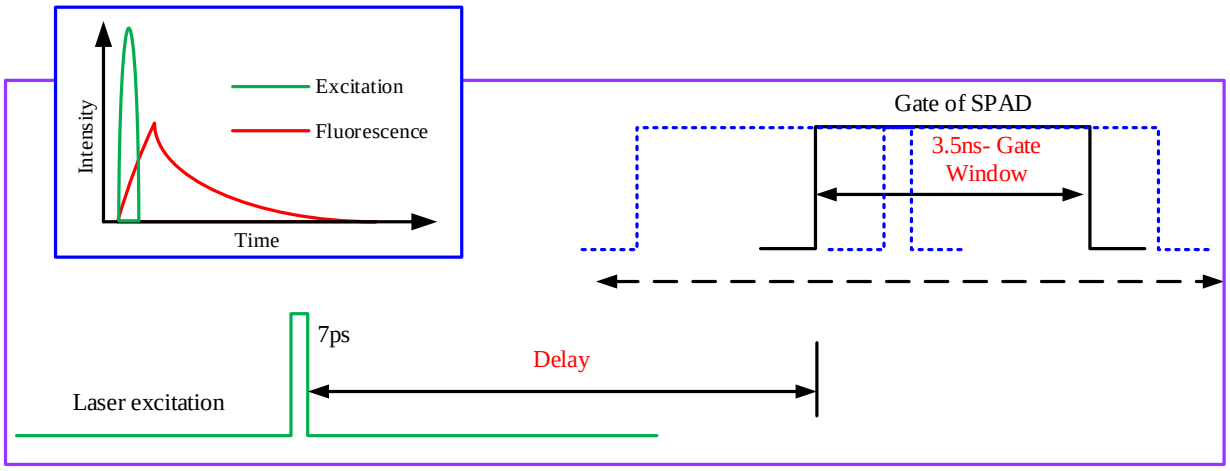


Figure 5.10: Principle of the time-gated fluorescence lifetime measurement

With the purpose of testing the feasibility of the TG-SPAD for fluorescence lifetime characterization, different fluorescence dyes were prepared and the fluorescence decay curves of the dyes were measured. Considering the excitation wavelength (532nm) of the laser in the setup, two types of fluorescence dyes were selected—Rhodamine B and Rhodamine 6G. When dissolved in water, the excitation and emission peaks of Rhodamine 6G are 525nm and 555nm respectively [163]. Its fluorescence lifetime was reported to be 4.08ns. The samples were purchased from Sigma-Aldrich.

Figure 5.11 shows the measured fluorescence decay curve of Rhodamine B and Rhodamine 6G, in which both the rising edge and exponential decay of the fluorescence are well displayed. From exponential fits to the curves, the extracted fluorescence lifetimes of Rhodamine B and Rhodamine 6G were 1.52ns and 3.94ns respectively. These results are slightly lower, but close to the reference values (1.68ns and 4.08ns), proving the feasibility of this time-gated setup for measurement of fluorescence decay.

5.3.2 Analysis of Fluorescence Decay Measurement and Discussions

Deviation between the measurements and the reference is caused by the timing response of the gate window to the incident photons, which is also related to the ‘pile-up’ effect, defined below.

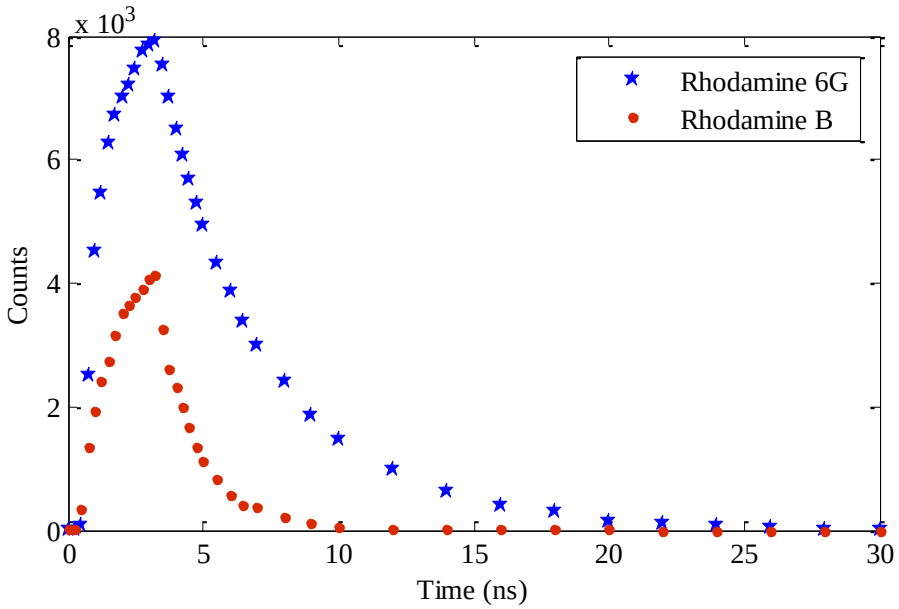


Figure 5.11: Fluorescence decays measurements of Rhodamine 6G and Rhodamine

In time-resolved measurements, after sample excitation by a pulsed laser, fast detection is used to count the number of photons emitted and record the emission time of each photon. In the case of a strong excitation, there are perhaps several photons emitted after an excitation, but at different times. However, limited by the deadtime of the SPAD, only the first arriving photon can be recorded. Thus, the time-resolved measurement can underestimate the signal, and this is known as the 'pile-up' effect [167].

The 'pile-up' effect has a strong impact on the TCSPC technique. It limits the maximum counting rate of the measurement. To minimize the 'pile-up' effect, usually the incident light is attenuated to control the number of photons emitted during each cycle (~1 photon). The time-gated measurement is less affected by the 'pile-up' effect. This is because different time delays are used (Δt), which ensures that photons emitted at different Δt have the same probability to be recorded. However, the 'pile-up' effect still exists within each gate window, especially when the measured signal is strong.

To investigate the 'pile-up' effect, not only the total photon number, but also the arrival time of each photon within the gate window was recorded by the high-speed, real-time oscilloscope (LeCroy SDA 18000 Serial Data Analyzer). Figure 5.12 shows the photon arrival time distributions within several gate windows at different time delays from the excitation laser pulse to the gate window. The delay time and the total photon number detected at this delay are given in the title of

each subplot. Since this was measured at the fluorescence decay curve, the number of counts decreased when increasing the delay time.

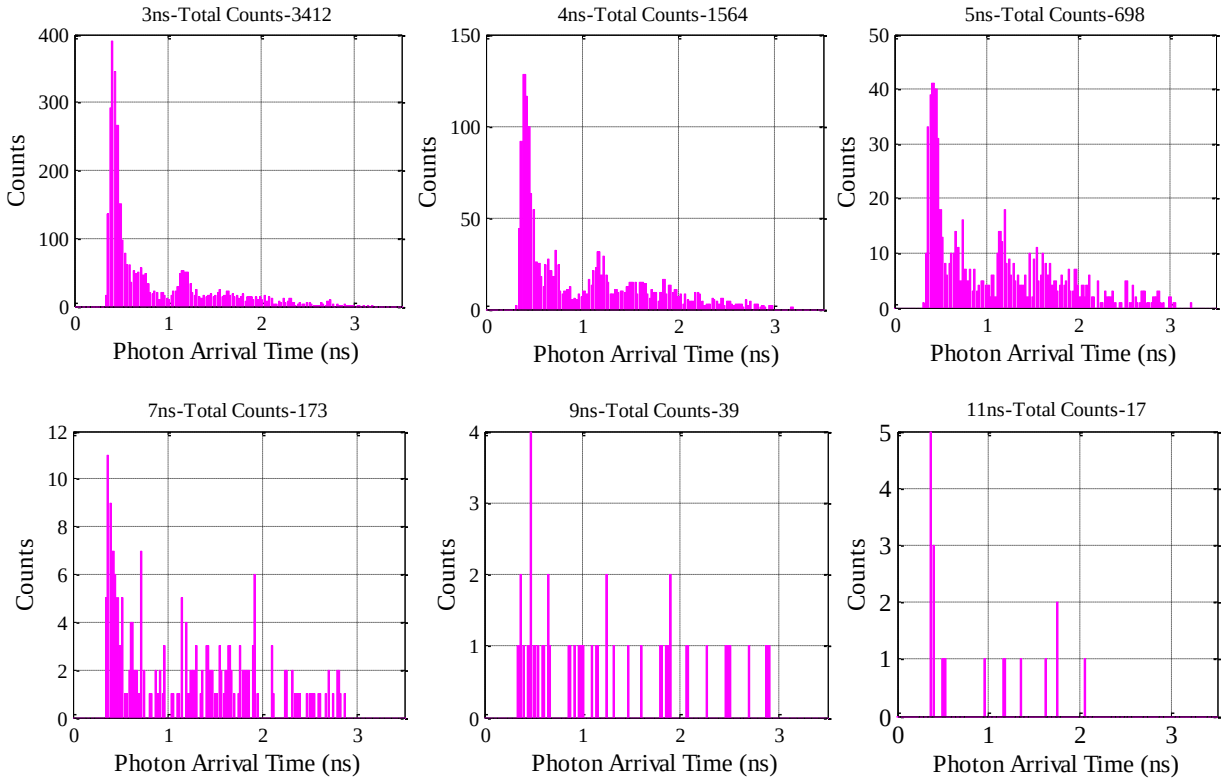


Figure 5.12: Photon arrival time distributions within each gate window of different time delays

A strong peak can be observed at the beginning of each gate window, and the number of counts decays quickly toward the end of the gate window. The decay is more pronounced in windows with shorter delays, since the fluorescence signal decays with time and when the fluorescence signal is strong, among the 10000 acquisitions of each delay, most detection events were triggered by photons arrived at the beginning of the gate window. However, the TG-SPAD can only record the first arrived photon, which reduces the detection probability of photons arriving later than the first arrived photon, such as those arriving by the end of the gate window. Therefore, the number of photons recorded is less than the real number of photons emitted, which means that photons are missed during detection. With the increasing of delay time and further decay of fluorescence signal, the photon missing situation is improved, and this phenomenon is similar to the 'pile-up' effect in TCSPC measurement.

Considering the "photon missing" problem, there was a calculation error when the fluorescence lifetime was extracted directly from the measured data. An extra correction is required to improve the measurement accuracy. The basic idea is to minimize the 'pile-up' effect existing in gate

window. Since for each gate window, photons emitted later may not be recorded by the TG-SPAD, because the detector is already triggered if there are photons arriving earlier. Therefore, usually a narrow gate window in the range of hundreds picoseconds is used to reduce the 'pile-up' effect. However, the gate window of this design is fixed by the on-chip pulse generators, for purpose of simplifying the synchronization setup during measurement. Therefore, 'pile-up' effect is unavoidable in this setup.

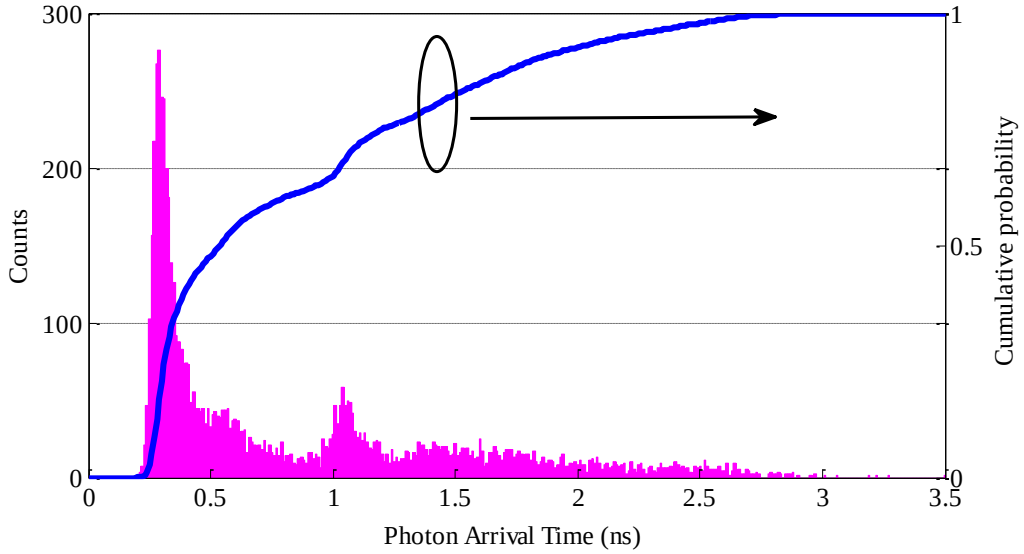


Figure 5.13: Histogram and cumulative probability of photons measured with a delay time of 4ns after the laser pulse excitation-Rhodamine 6G

Figure 5.13 shows the histogram and the cumulative probability of the photon counts with a delay time of 4ns after the laser pulse excitation. From the photon arrival time distribution we can see that photons arriving earlier in the gate window are numerous and the number of photons in the beginning of the gate window better represent the real number of photons arrived. To reduce the effect from photons missing in detection, data can be selected only from the beginning portion of the gate window by setting a cumulative probability. A compensation is also introduced to calculate the equivalent photon arrival time by weight of photon numbers.

$$t_n = \text{Delay} + \frac{\sum_{i=1}^{n} t_i \cdot c_i}{\text{C}_n},\ \text{C}_n = \sum_{i=1}^{n} c_i \tag{5.3}$$

$$\text{Cumulative probability } P(n) = \frac{C_n}{C_{Total}}$$

t_i : Photon arrival time within a gate window

c_i : Number of photons detected at t_i

C_{Total}: Total number of photon detected in a gate window

Eq. (5.3) gives the modified expression of photon emission time, where n is determined by the selected cumulative probabilities. With the modified photon number C_n and photon arrival time t_n on basis of the measured data, fluorescence lifetime for a selected cumulative probability is then extracted through the exponential fitting.

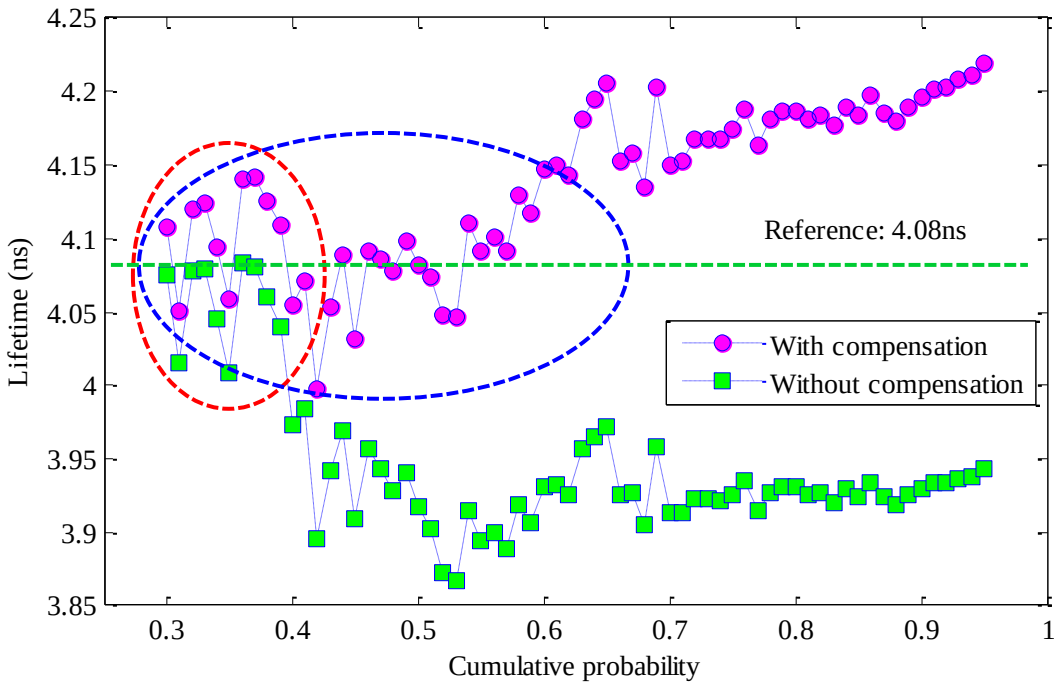


Figure 5.14: Fluorescence lifetime extracted with different cumulative probabilities of photons (Rhodamine 6G)

To characterize its dependence on the cumulative probability, the fluorescence lifetime was extracted with [Eq. (5.3)] and without photon arrival time compensation, as shown in Figure 5.14. Since cumulative probability determines the range of data picked from the gate window, a small cumulative probability refers to the early portion of the gate window, and cumulative probability of unity is corresponds to the entire gate window. The lifetime of Rhodamine 6G is reported to be 4.08ns (green dashed line). From the extraction results, lifetimes extracted from the early gate window are closer to the real fluorescence lifetime. The average fluorescence lifetimes for lower cumulative probabilities (red dashed ellipse) are 4.04ns and 4.10ns respectively for data with and without photon arrival time compensation, more accurate than the result (3.94ns) extracted from the entire gate window. This verifies the discussion above. In addition, with the increasing

cumulative probability, both results start to deviate from the reference value, indicating the influence of the 'pile-up' effect.

Comparing the two sets of data in Figure 5.14, the extracted lifetime with photon arrival time compensation is higher than those without compensation. This is because for longer time delays, the photon density is lower, and detection is less affected by the 'pile up' effect as that of short time delays. With a fixed cumulatively probability, t_n in Eq. (5.3) will shift to larger values for longer time delays. This shift slows down the decay of the processed fluorescence signal (C_n) and increases the extracted fluorescence lifetime. However, this compensation is useful because it provides stable data in a larger scale (blue dashed ellipse) of cumulative probability than the extraction without compensation (red dashed ellipse).

5.4 Future Improvements

A miniaturized time-gated spectrometer prototype was verified as feasible for applications in Raman spectroscopy and fluorescence lifetime measurement. The results from the verification are promising, confirming concepts and designs, and it also suggested directions for improvements in future development of time-gated spectrometers.

First, the acquisition time of a Raman spectrum is long, when using a single pixel detector to acquire data of a wavelength band. The mechanical adjustment of the detector on the translation stage was manual and slow. Much faster measurements can be achieved by using a motorized mechanical adjustment, but the ultimate speed up of the measurement, and perhaps the lowest cost, will be achieved with an array of TG-SPADs. Therefore, a future design of an array of TG-SPAD is recommended.

Second, there are critical trade-offs between spectral resolution and sensitivity in a miniaturized time-gated spectrometer. Higher spectral resolution requires small diameter fiber for better focusing on the Rowland circle of the concave grating, while higher sensitivity requires a larger diameter fiber for better optical coupling of the optical signal. An optimum was obtained by limiting the Raman spectra measurement only to the dominant peaks and the time-gated photodetection allows for suppression of the fluorescence signal. Future work should determine the rules for solving the trade-off in manufacturing miniaturized spectrometers.

Third, while this work achieved miniaturization and cost reduction of the wavelength selector and the time-gated single photon detector, more work is required to integrate the control and signal

processing on chip. The control and acquisition blocks have to include delay lines, pulse width analyzers and counters. To reduce the chip area, some digital circuits can be replaced with analog, e.g., pulse counters for Raman spectroscopy can use analog counting.

Fourth, an important achievement in the proposed system is the management of a fixed gate window to be used as variable width *detection window*. The fixed gate window allowed for the on-chip integration of the control of the TG-SPAD by means of pulse generators, triggered by a single signal. The method for management of the variable width *detection window* is also simple, requiring only a delay unit. In this chapter, these methods were proven to be suitable to suppress fluorescence during Raman measurements or to measure the fluorescence decay. Future work should consider integrating the delay unit on chip, which will create a multifunctional TG-SPAD font-end feasible for various applications of Raman and fluorescence measurements. A good candidate for the delay unit is the voltage control delay line (VCDL) that is commonly used in TCSPC systems.

5.5 Summary

In this chapter, a prototype of the miniaturized time-gated spectrometer was built with a concave grating and a TG-SPAD, both fabricated by inexpensive technologies. To carry out the time-resolved measurements, the system was optically synchronized to the laser pulse excitation by a beam splitter and a high speed photodiode. Timing resolution of the time-gated system was tested with a 7ps pulsed laser, and better than 60ps timing resolution was achieved. The time-gated spectrometer was used for both Raman spectrum and fluorescence lifetime measurements. Suppression of the fluorescence in a Raman spectrum measurement was achieved by narrowing the *detection window*, and the major Raman peaks of the sample were resolved with this low-cost system, while a commercial spectrometer operating in the free running mode could not resolve any Raman signal from the strong fluorescence background. Fluorescence lifetimes of Rhodamine B and Rhodamine 6G were extracted from the time-gated measurements. The extracted values of the fluorescence lifetimes were very close to reference data from the literature. During the verification tests of the time-gated spectrometer, issues such as timing jitter in the triggering of the pulsed laser were solved. Recommendations for future improvements of the time-gated spectrometer, based on comprehensive analyses of the verification tests, were provided.

Chapter 6

Conclusions and Recommendations for Future Work

In this thesis, a low-cost time-gated spectrometer was designed and a prototype was fabricated. This chapter summarizes the research work and provides recommendations for future work.

6.1 Conclusions

The purposes of this work were to prove concepts and build a compact, low-cost spectrometer towards field applications. Raman spectroscopy was selected because of its non-contact and non-destructive properties. In addition, since water is a weak Raman scatterer, Raman spectroscopy is suitable for direct measurement of water samples. However, to measure the weak Raman spectra, commercial bench-top Raman spectrometers are very expensive and bulky, especially when strong fluorescence signal is present. Portable Raman spectrometers are now commercially available. These Raman spectrometers are small in size and easy to use, but still expensive, especially those with fluorescence rejection. Thus, the main challenge of this work is the design of a compact, low-cost Raman spectrometer with fluorescence suppression capability.

A Raman spectrometer contains four basic components—an excitation source, light illumination and collection optics, a wavelength selector, and a detector. This work focused on the wavelength selector and the detector. Addressing system miniaturization, a concave grating was used for wavelength selection. An important advantage of a concave grating is that it can perform both functions of light wavelength separation and focusing without the need for collimating and focusing mirrors. Therefore, a concave grating based system is more compact than a system with a planar grating.

Grating theory was reviewed, and the design of concave grating was presented. A commercial simulator was used to investigate the diffraction efficiency, through which several design parameters were determined, such as the range of incident angle, grating constant, and the coating material. The ranges of the design parameters were further narrowed through simulations of the spectral resolution. Based on aberration theory, a simplified algorithm was proposed to calculate the spectral resolution of a concave grating based system. A flat-field concave grating was also designed to provide a linear horizontal focal curve. In order to reduce the cost, the concave grating was fabricated by a custom holographic method, on the surface of plano-concave lenses. The grating profile was measured by AFM and SEM, followed by the characterizations of the dispersion and focusing properties of the grating. Spectral resolution of the concave grating was tested, and good agreement with the result calculated by the proposed algorithm was obtained.

Due to the low Raman scattering efficiency, detection of a Raman signal is very challenging, especially when a strong fluorescence background is present. Considering the different temporal distributions of excitation, Raman scattering and fluorescence emission, the time-gated detection mechanism was chosen to measure the Raman spectrum. In this detection strategy, the detection of Raman photons is in a short time window after the pulse excitation, when the fluorescence signal is still not present or very weak. A time-gated single photon avalanche diode (TG-SPAD) front-end was designed. To reduce the cost of the TG-SPAD, the front-end was implemented in a standard CMOS technology.

The chips were fabricated in the 130nm CMOS technology of IBM. Avalanche diodes were first characterized to extract relevant parameters of the diode when the diode works in the avalanche breakdown regime, which are of great importance for fast quenching and resetting of the SPAD. On-chip pulse generators were designed and included in the TG-SPAD, so that all control signals are generated on-chip by simple triggering from an external signal for precise synchronization.

The TG-SPAD front-end is with a fixed gate window of 3.5ns, determined by the design of on-chip pulse generators. The photon arrival time within each gate window can be obtained from the pulse width of the pixel's output signal. This TG-SPAD front-end requires only 5 transistors, aiming at improving the pixel fill factor and reducing the parasitic capacitance of the SPAD. The pixel functionality was verified under illumination with a halogen lamp, and random photon arrival times were obtained. The performance characterization of the TG-SPAD front-end was then carried out, including the dark count probability per gate window (DCP_{GW}), afterpulsing probability, and

photon detection efficiency (PDE). The temperature dependence of DCP_{GW} was investigated. From the Arrhenius plot of DCP_{GW}, a low activation energy was extracted. This low value indicates an increased density of non-mid-gap traps and contribution from tunneling for SPADs fabricated by DSM CMOS technology. The front-end operated well at gating frequencies (f_G) up to 100MHz. By applying a hold-off time >16ns, ($f_G \leq 50$MHz), the afterpulsing probability is low (<1%) and negligible. However, because of the top passivation layers of this inexpensive CMOS process, a relatively low PDE ~1%-3% was measured.

Combining the concave grating and the TG-SPAD with a commercial pulsed green laser (532nm, 7ps), the prototype of a miniaturized time-gated spectrometer was investigated. Since a single pixel TG-SPAD was used, then the TG-SPAD was mounted on a translation stage to acquire the Raman spectrum at different wavelengths. Characterization of the time-gated spectrometer indicated that better than 60ps temporal resolution was achieved. In order to test the efficiency of the time-gated spectrometer for fluorescence suppression, the Raman spectra of a fluorescence dye (Rhodamine B) were measured. This dye emits a strong fluorescence signal. In addition, the fluorescence lifetime of Rhodamine B is short, ~1.7ns, challenging Raman photon acquisition with the 3.5ns gate window. Therefore, a dedicated method to shorten the gate window into sub-ns *detection window* was developed. Using this technique, *detection windows* down to 250ps were achieved, and verified by the small variation of the Rayleigh scattering spectra at the wavelength of the laser excitation. By reducing the *detection window* from 3ns to 250ps, the strong fluorescence background was suppressed in the acquisition of the Raman spectrum, proving the efficiency of time gating for fluorescence suppression. The reduced *detection window* also allowed for resolving two Raman peaks of the fluorescence dye. These Raman peaks were not resolved with a commercial spectrometer operated in free-running mode. The Raman shifts of the two peaks were stable for different *detection windows*, and the shifts were close to published data measured by a commercial surface enhanced Raman spectrometer. Therefore, it was concluded that a compact and inexpensive Raman spectrometer for field applications is an achievable goal in the near future.

To extend the function of the time-gated spectrometer, the time-resolved measurement of fluorescence lifetime was carried out. An experimental verification of the feasibility of the TG-SPAD front-end for fluorescence lifetime measurement was presented for two types of fluorescence dyes. The extracted fluorescence lifetimes are very close to reference data from literature.

6.2 Future Work

In this thesis, a low-cost time-gated Raman spectrometer was designed and a prototype was built. The verification tests of the prototype system were carried out, resolving major Raman peaks of a fluorescent dye. However, owing to the low spectral resolution, the weak Raman peaks were not resolved. The main drawbacks of the current system are the poor spectral resolution, low photon detection efficiency, and complex synchronization setup. To extend its application, improvements in the following aspects are necessary:

1) *Spectral resolution*: The poor spectral resolution of the miniaturized spectrometer is caused by several factors. First, the concave grating used in this setup has a constant line space. In the Rowland configuration, the horizontal focal curve of the concave grating overlaps with the Rowland circle. To date, no CMOS detector with curved surface is known. In this situation, using a planar detector at the curved Rowland circle of a concave grating degrades the spectral resolution. Therefore, a flat-field concave grating was designed and described in chapter 3, but the varied line spacing concave grating is usually fabricated by expensive processes, in which grating grooves are formed individually. Alternatively, the varied line space concave grating can also be fabricated by the holographic method with two laser sources. Grooves fabricated by this method are formed simultaneously, so the process is fast and the cost is relatively lower. Therefore, research on cost-effective formation of concave gratings with variable line spacing would be beneficial for improving the spectral resolution of the system. Second, spectral resolution is a function of several design parameters, including the incident angle, grating constant, grating radius, and entrance slit width. To gain high enough coupling efficiency, a multimode fiber with 200μm diameter was used in this system for light transmission. The large fiber diameter broadens the spectra and degrades the spectral resolution. To improve the resolution, a single mode fiber can be used, but extra optics is required to improve the coupling efficiency. In addition, the concave grating can be fabricated on a substrate with larger radius, but this will increase the size of the overall system.
2) *Photon detection efficiency*: PDE of the SPAD is related to the transmission coefficient of photons in the passivation layers, absorption coefficient of the material, and the triggering probability of an avalanche. The absorption coefficient depends on the wavelength of the incident photon and the material, both of which cannot be modified in a standard CMOS

process. The triggering probability can be increased by applying higher excess biases, but it also raises the dark count probability. One solution is to use a different CMOS technology with wider depletion region, such as a high-voltage (HV) process (0.35μm). Otherwise, it is not desirable to increase the PDE of the SPAD by simply increasing the excess bias. If a standard CMOS process is used, then a practical way to improve the PDE is to increase transmission probability of photons to the absorption region of the SPAD. This can be realized by removing the top passivation layers of the SPAD by post processing, or using a micro-lens array to make the incident light more focused when it reaches the active region of the SPAD, which eventually reduces surface reflections. However, any of these approaches require individual processing of chips, which will raise the cost of the TG-SPAD.

3) *Synchronization, timing, and acquisition*: Attention must be paid for synchronization in time-gated spectrometers. To synchronize a pulsed laser and a nanosecond gate window, global synchronization of the system is very challenging. Instead, a local optical synchronization of the TG-SPAD to a laser pulse is preferred, since it solves the problem of laser triggering jitter. In addition, a high-speed oscilloscope was used to count the number of photons arrived in a given period, and such an oscilloscope is neither small in size nor inexpensive. Future developments of time-gated spectrometers should use simpler on-chip modules for synchronization, management of delay, pulse counting, and pulse width analysis. Optical triggering of the time-gated spectrometer was solved by on-chip integration of pulse generators that are triggered by a photodiode. It is also necessary to integrate the delay unit and the processing of the pulse from the TG-SPAD in order to have a compact spectrometer. A good candidate for integration of the delay unit is a voltage controlled delay line, made of current-starved CMOS inverters. Regarding the acquisition, the high speed oscilloscope can be replaced with on-chip counter, either digital or analog. There should also be a pulse-width analyzer, to evaluate the adjustment of delays and for fluorescence lifetime measurement. Lastly, but not the least, a TG-SPAD array (ideally, with simultaneous counting from all pixels) has to be designed to accelerate the spectra acquisition.

The work presented in this thesis was a pilot research and development project, targeting a compact and low-cost Raman spectrometer for field applications. It is shown that the system

miniaturization is achievable with a concave grating based wavelength selector, and by designing a TG-SPAD front-end. The fabrication of the TG-SPAD front-end in a mainstream 130nm CMOS technology has significantly reduced the cost of the spectrometer. Performance tests have provided for feasibility, challenges, and solutions in building miniaturized Raman spectrometer for field application.

Appendix I: Design of Flat-Field Concave Gratings

The slope between Q and Q_0 of arbitrary wavelength (λ) and reference wavelength (λ_0) can be written as,

$$k_{QQ_0}(t)=\frac{r_{b0}\cos\beta_0-r_b\cos\beta}{r_b\sin\beta-r_{b0}\sin\beta_0}=\frac{r_{b0}\sqrt{1-(mG_0\lambda_0-\sin\alpha)^2}-r_b\sqrt{1-t^2}}{r_b\cdot t-r_{b0}(mG_0\lambda_0-\sin\alpha)}, \tag{I.1}$$

$$\text{where } r_b=\frac{\cos^2\beta}{\frac{\cos\beta}{R}-mG_0\lambda\cdot H_{20}}=\frac{R\left(1-t^2\right)}{\left(1-t^2\right)^{\frac{1}{2}}-R\cdot t\cdot H_{20}-R\sin\alpha\cdot H_{20}},\ t=mG_0\lambda-\sin\alpha\ .$$

The power series expansion of Eq. (I.1) in terms of t, gives

$$k_{QQ_0}(t)=g(t)=g_0+g'(0)t+\frac{1}{2}g''(0)t^2+\frac{1}{6}g'''(0)t^3+... \tag{I.2}$$

To guarantee a constant slope, which means keeping only the g_0 term, then the sum of all other terms is set to zero. Since |t|<1, high order g terms become small and their contributions to the slope are negligible. Therefore, to make the slope constant, the problem is to solve for g'(0)=0.

Eq. (I.1) is rewritten as

$$k_{QQ_0}(t)=\frac{A-r_b\sqrt{1-t^2}}{r_b\cdot t-B}\ , \tag{I.3}$$

$$\text{where: } A=r_{b0}\sqrt{1-(mG_0\lambda_0-\sin\alpha)^2},\ B=r_{b0}(mG_0\lambda_0-\sin\alpha)\ .$$

$$g'(0)=\left(\frac{A-r_b\sqrt{1-t^2}}{r_b\cdot t-B}\right)'_{t=0}=\left(\frac{A}{r_b\cdot t-B}\right)'_{t=0}-\left(\frac{r_b\sqrt{1-t^2}}{r_b\cdot t-B}\right)'_{t=0} \tag{I.4}$$

$$\left(\frac{A}{r_b\cdot t-B}\right)'_{t=0}=\left(\frac{-A\left(r_b+t\cdot r_b'\right)}{\left(r_b\cdot t-B\right)^2}\right)_{t=0}=\left(\frac{-Ar_b}{B^2}\right)_{t=0}. \tag{I.5}$$

$$\left(\frac{r_b\sqrt{1-t^2}}{r_b\cdot t-B}\right)'_{t=0}=\left(\frac{r_b'\sqrt{1-t^2}\cdot\left(r_b\cdot t-B\right)+r_b\sqrt{1-t^2}'\cdot\left(r_b\cdot t-B\right)-r_b\sqrt{1-t^2}r_b}{\left(r_b\cdot t-B\right)^2}\right)_{t=0}=\left(\frac{-Br_b'-r_b^2}{B^2}\right)_{t=0} \tag{I.6}$$

$$r_b(t=0)=\frac{R}{1-R\sin\alpha\cdot H_{20}},\ r_b'(t=0)=\frac{R^2H_{20}}{\left(1-R\sin\alpha\cdot H_{20}\right)^2} \tag{I.7}$$

Substituting Eqs. (I.5-7) to Eq. (I.4) and setting g'(0) to zero gives

$$H_{20}=\frac{\sqrt{1-t_0^2}}{RmG_0\lambda_0}=\frac{\sqrt{1-\left(mG_0\lambda_0-\sin\alpha\right)^2}}{RmG_0\lambda_0} \tag{I.8}$$

Appendix II: Challenges and Solutions for the Fabrication of Concave Gratings

This Appendix contains a supplementary record on several attempts for fabrication of concave grating. These attempts allowed for the optimization of the fabrication conditions. The final fabrication approach is presented in Chapter 3.

To test the efficiency of the holographic setup shown in Figure 3.20(b), fabrication of a planar grating on a silicon substrate was tried. Table II.1 lists the process parameters. The fabrication process flow is as follows. The photoresist S1808 was diluted with a thinner with dilution ratio of 2:3. The silicon substrate was then spin-coated with the thinned solution at speed of 5000rpm for 30s, and a uniformly thin layer of ~200nm thickness was achieved. The silicon substrate was then soft baked on a hot plate at temperature of 80°C for 2mins. The baked silicon substrate was mounted on the sample holder of the holographic setup. The laser shuttle was opened for ~40s to expose the photoresist to the interference pattern. The sample was finally developed in the solution CD30 for a period of ~1min 30s. The development time was controlled by comparing the color contrast between the exposed and an unexposed region of the silicon substrate. During the development, a rainbow was observed to appear from the sample in the developer. The rainbow indicated that diffraction structures were successfully formed on the silicon substrate. The intensity of the rainbow is also a good way to control the development time.

Table II.1 Grating fabrication parameters

Parameter	Value
Laser wavelength	325nm
Laser power	200μW
Photoresist (PR)	S1808
Solution	PR (2) : Thinner (3)
Spinner	5000rpm, 30s
Exposure time	40s
Developer	CD30, 1min 30s

The parameters listed in Table II.1 are only reference values for a specific condition. In case of different conditions, these parameters must be adjusted according to different type of substrates, thickness of photoresist, concentration of developer, and power of the laser.

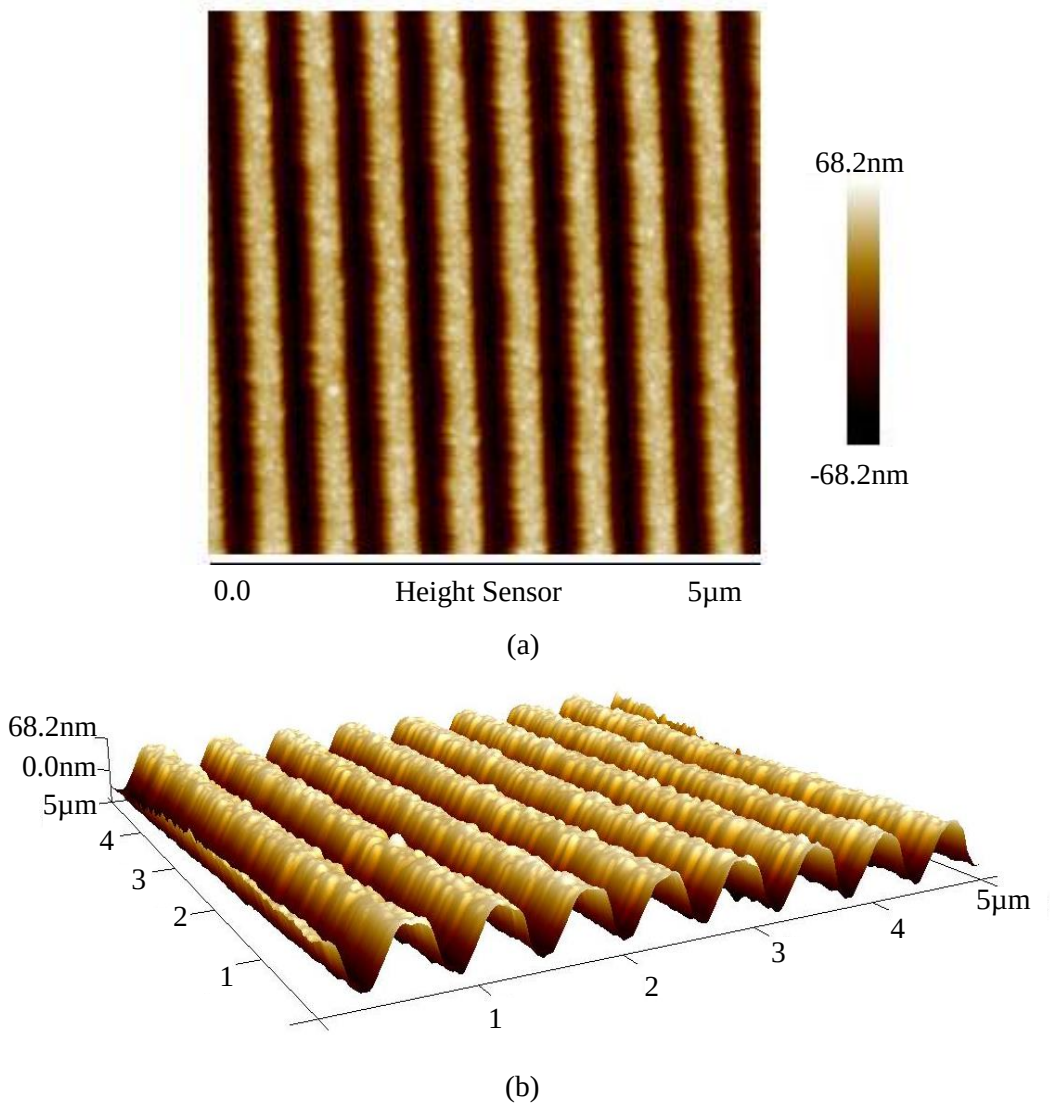


Figure II.1: AFM image of a grating on a silicon substrate: (a) 2D surface image; (b) 3D image

After fabrication, the surface profile of the grating was characterized. The surface was scanned with an AFM. Figure II.1 shows the 2D and 3D images of the grating structures. Periodic grooves are seen, which proved the efficiency of the fabrication method for producing grating structures. The grating constant is ~580nm, and the depth of each groove is ~136nm. The sinusoidal groove profile can be observed from the 3D image. If a blazed grating is preferred, the sample can be further processed by etching the silicon to form groves with a triangular profile. The photoresist in

this case will serve as an etching mask. The final step of the grating fabrication is to coat the surface with a thin layer of Au, because the photoresist is not reflective.

Fabrication of a concave grating might be different from that of a planar grating. The first challenge comes from the substrate, which cannot be a silicon wafer. The silicon-based concave grating reported in [52] was a cylindrical concave grating, rather than a spherical concave grating, and it was fabricated by the costly deep x-ray lithography. On the other hand, the most common concave substrate is glass with a spherical surface. To test the feasibility to form grating structures on a concave glass substrate, a double concave watch glass was tried first.

(a) (b)

Figure II.2: (a) Double concave watch glass sample; (b) Glass sample after fabrication

Figure II.2 (a) shows the image of the double concave watch glass sample. This sample was purchased from the chemical store of McMaster University ($3), and had a diameter of ~3cm. To produce grating structures on a concave glass substrate, there are several uncertainties. First, no appropriate chuck can be used to hold a concave surface during the spin coating with photoresist. Second, when using the convex surface, it is difficult to guarantee a uniform coating, since the photoresist may flow from the top to the bottom of the surface. Third, soft bake of the concave substrate using a hot plate cannot guarantee uniform heating of the entire surface, since only the circular edge touches the hot plate. Fourth, the development time is difficult to control, and because the glass is transparent, there is not a clear color contrast between exposed and unexposed areas.

To solve the first problem, the glass substrate was glued on a flat glass slide using a strong double side tape. Now the glass substrate can spin with the glass slide. Since the adhesive force of the tape is not as strong as the vacuum, and the glass substrate is heavier than the glass slide, there is risk for the glass substrate to fly off the glass slide if a high spin speed is used. Spin speed and

time determine the thickness and uniformity of the coating, and a high spin speed is preferred for a thin layer of coating. To determine the maximum achievable spin rate, the spin rate was adjusted from 500rpm to 1000rpm. A short spin time of 10s was used to avoid the glass substrate flying off the glass slide.

After the coating of photoresist, different soft bake times and temperatures were tried for both hot plate and oven for a uniform baking. Unfortunately, due to the thick photoresist layer, down flowing of the photoresist after spin coating, and non-uniform baking, an accumulation of photoresist was observed on the convex surface. This accumulation is shown in Figure II.2 (b).

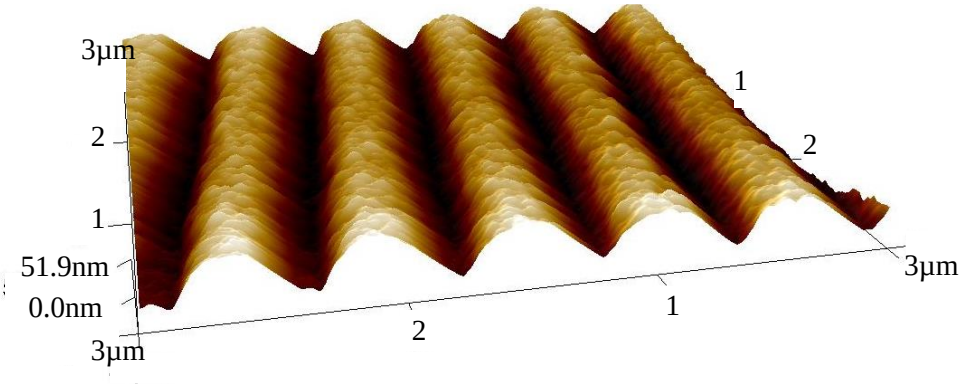


Figure II.3: 3D AFM image of the grating on concave glass substrate (measured at the top of the substrate)

Considering the transparency of the watch glass substrate, the sample after exposure was developed for different times from 30s to 1min30s with a step of 15s. Relatively strong diffraction was observed when the development time was about 1min. Figure II.2 (b) shows the image of the concave glass after the holographic fabrication. A rainbow appeared on the surface, indicating that the grating structure was successfully formed on the convex surface of the watch glass. In order to characterize the surface profile after fabrication, the surface was measured using an AFM. However, owing to the curved surface, only a narrow region of 3µm x 3µm on top of the convex side was scanned. The 3D AFM image shown in Figure II.3 proved that grating structures have been fabricated on the curved surface of the watch glass.

Although diffraction structures could be produced on the convex side of a watch glass, fabrication of a grating on the concave side failed owing to challenges in the spin-coating with photoresist. Moreover, the surface of a watch glass is not guaranteed to be perfectly spherical. These problems were solved by using a plano-concave lens, as explained in chapter 3.

References

1. P.H. Gleick, *Water in crisis: a guide to the world's fresh water resources.*, Oxford University Press, Inc., (1993).
2. Global Water Partnership. *http://www.gwp.org/Press-Room/Water-Statistics/* [Online].
3. World Water Council. *http://www.worldwatercouncil.org/library/archives/water-crisis/* [Online].
4. J. Bartram and R. Ballance, *Water quality monitoring: a practical guide to the design and implementation of freshwater quality studies and monitoring programmes*, CRC Press, (1996).
5. A. Gowen, R. Tsenkova, M. Bruen and C. O'donnell, "Vibrational spectroscopy for analysis of water for human use and in aquatic ecosystems," *Crit.Rev.Environ.Sci.Technol.*, vol. 42**(23)**, pp. 2546-2573, (2012).
6. United Nations Environment Programme. *http://www.unep.org/gemswater/GlobalNetwork/tabid/78238/Default.aspx* [Online].
7. J. Hasan, D. Goldbloom-Helzner, A. Ichida, T. Rouse and M. Gibson, "Technologies and techniques for early warning systems to monitor and evaluate drinking water quality: A state-of-the-art review," *Technologies and techniques for early warning systems to monitor and evaluate drinking water quality: A state-of-the-art review*, (2005).
8. Z. Li, M.J. Deen, S. Kumar and P.R. Selvaganapathy, "Raman spectroscopy for in-line water quality Monitoring—Instrumentation and potential," *Sensors*, vol. 14**(9)**, pp. 17275-17303, (2014).
9. A.L. Givan, *Flow cytometry: first principles*, John Wiley & Sons, (2013).
10. H.M. Shapiro, *Practical flow cytometry*, John Wiley & Sons, (2005).
11. S. Ahuja and S. Scypinski, *Handbook of modern pharmaceutical analysis*, Academic press, (2010).
12. H.M. McNair and J.M. Miller, *Basic gas chromatography*, John Wiley & Sons, (2011).
13. M.R. Siddiqui, Z.A. AlOthman and N. Rahman, "Analytical techniques in pharmaceutical analysis: A review," *Arabian Journal of Chemistry*, (2013).
14. M. Harz, P. Rösch and J. Popp, "Vibrational spectroscopy—A powerful tool for the rapid identification of microbial cells at the single-cell level," *Cytometry Part A*, vol. 75**(2)**, pp. 104-113, (2009).
15. Luisa Andronie, "The adsorption behaviour of buffered aspirin monitored by raman and surface-enhanced raman spectroscopy," *International Journal of the Bioflux Society*, vol. 6**(4)**, pp. 187-190, (2014).
16. E. Smith and G. Dent, *Modern Raman spectroscopy: a practical approach*, Wiley, (2005).
17. W. Lin and Z. Li, "Detection and quantification of trace organic contaminants in water using the FT-IR-attenuated total reflectance technique," *Anal.Chem.*, vol. 82**(2)**, pp. 505-515, (2009).
18. D. Pérez-Quintanilla, A. Sánchez, I. del Hierro, M. Fajardo and I. Sierra, "Preconcentration of zn(II) in water samples using a new hybrid SBA-15-based material," *J.Hazard.Mater.*, vol. 166**(2–3)**, pp. 1449-1458, (2009).
19. G. Reich, "Near-infrared spectroscopy and imaging: Basic principles and pharmaceutical applications," *Adv.Drug Deliv.Rev.*, vol. 57**(8)**, pp. 1109-1143, (2005).
20. V.J. Frost and K. Molt, "Analysis of aqueous solutions by near-infrared spectrometry (NIRS) III. binary mixtures of inorganic salts in water," *J.Mol.Struct.*, vol. 410–411**(0)**, pp. 573-579, (1997).
21. A. Sakudo, R. Tsenkova, K. Tei, T. Onozuka, K. Ikuta, E. Yoshimura and T. Onodera, "Comparison of the vibration mode of metals in HNO3 by a partial least-squares regression analysis of near-infrared spectra," *Biosci.Biotechnol.Biochem.*, vol. 70**(7)**, pp. 1578-1583, (2006).
22. S. Keren, C. Zavaleta, Z. Cheng, A. de la Zerda, O. Gheysens and S.S. Gambhir, "Noninvasive molecular imaging of small living subjects using raman spectroscopy," *Proc.Natl.Acad.Sci.U.S.A.*, vol. 105**(15)**, pp. 5844-5849, (2008).

23. C.W. Freudiger, W. Min, B.G. Saar, S. Lu, G.R. Holtom, C. He, J.C. Tsai, J.X. Kang and X.S. Xie, "Label-free biomedical imaging with high sensitivity by stimulated raman scattering microscopy," *Science*, vol. 322**(5909)**, pp. 1857-1861, (2008).

24. J. Popp, C. Krafft and T. Mayerhöfer, "Modern raman spectroscopy for biomedical applications," *Optik & Photonik*, vol. 6**(4)**, pp. 24-28, (2011).

25. E.B. Bradley and C.A. Frenzel, "On the exploitation of laser raman spectroscopy for detection and identification of molecular water pollutants," *Water Res.*, vol. 4**(1)**, pp. 125-128, (1970).

26. T.W. Collette and T.L. Williams, "The role of raman spectroscopy in the analytical chemistry of potable water," *Journal of Environmental Monitoring*, vol. 4**(1)**, pp. 27-34, (2002).

27. D. Boas A, C. Pitris and N. Ramanujam, *Handbook of biomedical optics*, Boca Raton, FL: CRC press, (2011).

28. J.R. Lackowicz, *Principles of fluorescence spectroscopy*, New York: Springer Science Business Media, (2006).

29. M.Y. Berezin and S. Achilefu, "Fluorescence lifetime measurements and biological imaging," *Chem.Rev.*, vol. 110**(5)**, pp. 2641-2684, (2010).

30. Brain Icon. *http://www.wpclipart.com/medical/anatomy/brain/brain_icon.jpg.html* [Online].

31. A. Tosi, A.D. Mora, F. Zappa, A. Gulinatti, D. Contini, A. Pifferi, L. Spinelli, A. Torricelli and R. Cubeddu, "Fast-gated single-photon counting technique widens dynamic range and speeds up acquisition time in time-resolved measurements," *Optics Express*, vol. 19**(11)**, pp. 10735-10746, (2011).

32. H. Alhemsi, Zhiyun Li and M.J. Deen, "Time-resolved near-infrared spectroscopic imaging systems," in *Electronics, Communications and Photonics Conference (SIECPC), 2013 Saudi International*, 2013, pp. 1-6.

33. M.J. Deen and E. Thompson, "Design and simulated performance of a CARS spectrometer for dynamic temperature measurements using electronic heterodyning," *Appl.Opt.*, vol. 28**(7)**, pp. 1409-1416, (1989).

34. J.R. Ferraro, *Introductory Raman spectroscopy*, Academic press, (2003).

35. CHEMwiki. *http://chemwiki.ucdavis.edu/Physical_Chemistry/Spectroscopy/Electronic_Spectroscopy/Jablonski_diagram* [Online].

36. R.L. McCreery, *Raman spectroscopy for chemical analysis*, Wiley-Interscience, (2005).

37. H. Kogelnik and S. Porto, "Continuous helium-neon red laser as a raman source," *Journal Optics Society American*, vol. 53**(12)**, (1963).

38. T. Hirschfeld and B. Chase, "FT-raman spectroscopy: Development and justification," *Appl.Spectrosc.*, vol. 40**(2)**, pp. 133-137, (1986).

39. M. Fujiwara, H. Hamaguchi and M. Tasumi, "Measurements of spontaneous raman scattering with nd: YAG 1064-nm laser light," *Appl.Spectrosc.*, vol. 40**(2)**, pp. 137-139, (1986).

40. C. Belin, C. Quellec, M. Lamotte, M. Ewald and P. Simon, "Characterization by fluorescence of the dissolved organic matter in natural water. application to fractions obtained by tangential ultrafiltration and XAD resin isolation," *Environ.Technol.*, vol. 14**(12)**, pp. 1131-1144, (1993).

41. N. Faramarzpour, M.M. El-Desouki, M.J. Deen, S. Shirani and Q. Fang, "CMOS photodetector systems for low-level light applications," *J.Mater.Sci.: Mater.Electron.*, vol. 20**(1)**, pp. 87-93, (2009).

42. F.d.S. Campos, N. Faramarzpour, O. Marinov, M.J. Deen and J.W. Swart, "Photodetection with gate-controlled lateral BJTs from standard CMOS technology," *IEEE Sensors Journal*, pp. 1554-1563, (2013).

43. N. Faramarzpour, M.J. Deen, S. Shirani, Q. Fang, L. Liu, F. de Souza Campos and J.W. Swart, "CMOS-based active pixel for low-light-level detection: Analysis and measurements," *IEEE Trans. Electron Devices*, vol. 54**(12)**, pp. 3229-3237, (2007).

44. C. Xie, M.A. Dinno and Y. Li, "Near-infrared raman spectroscopy of single optically trapped biological cells," *Opt.Lett.*, vol. 27**(4)**, pp. 249-251, (2002).

45. M.D. Hargreaves, K. Page, T. Munshi, R. Tomsett, G. Lynch and H.G. Edwards, "Analysis of seized drugs using portable raman spectroscopy in an airport environment—a proof of principle study," *J.Raman Spectrosc.*, vol. 39**(7)**, pp. 873-880, (2008).

46. B. Yang, M.D. Morris and H. Owen, "Holographic notch filter for low-wavenumber stokes and anti-stokes raman spectroscopy," *Appl.Spectrosc.*, vol. 45**(9)**, pp. 1533-1536, (1991).

47. S. Grabarnik, R. Wolffenbuttel, A. Emadi, M. Loktev, E. Sokolova and G. Vdovin, "Planar double-grating microspectrometer," *Optics Express*, vol. 15**(6)**, pp. 3581-3588, (2007).

48. S. Grabarnik, A. Emadi, H. Wu, G. de Graaf and R.F. Wolffenbuttel, "High-resolution microspectrometer with an aberration-correcting planar grating," *Appl.Opt.*, vol. 47**(34)**, pp. 6442-6447, (2008).

49. Z. Li, M.J. Deen, Q. Fang and P. Selvaganapathy, "Design of a flat field concave-grating-based micro-raman spectrometer for environmental applications," *Appl.Opt.*, vol. 51**(28)**, pp. 6855-6863, (2012).

50. R. Brunner, M. Burkhardt, K. Rudolf and N. Correns, "Microspectrometer based on holographically recorded diffractive elements using supplementary holograms," *Optics Express*, vol. 16**(16)**, pp. 12239-12250, (2008).

51. Y. Chen, Y. Lee, J. Chang and L.A. Wang, "Fabrication of concave gratings by curved surface UV-nanoimprint lithography," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 26**(5)**, pp. 1690-1695, (2008).

52. C. Ko, B. Shew and S. Hsu, "Micrograting fabricated by deep x-ray lithography for optical communications," *Optical Engineering*, vol. 46**(4)**, pp. 048001-048001-7, (2007).

53. S. Grabarnik, A. Emadi, E. Sokolova, G. Vdovin and R.F. Wolffenbuttel, "Optimal implementation of a microspectrometer based on a single flat diffraction grating," *Appl.Opt.*, vol. 47**(12)**, pp. 2082-2090, (2008).

54. S. Grabarnik, A. Emadi, H. Wu, G. de Graaf and R. Wolffenbuttel, "Microspectrometer with a concave grating fabricated in a MEMS technology," *Procedia Chemistry*, vol. 1**(1)**, pp. 401-404, (2009).

55. C. Ko and M.R. Lee, "Design and fabrication of a microspectrometer based on silicon concave micrograting," *Optical Engineering*, vol. 50**(8)**, pp. 084401-084401-10, (2011).

56. G. Chantry and H. Gebbie, "Interferometric raman spectroscopy using infra-red excitation," *Nature*, vol. 203, pp. 1052-1053, (1964).

57. H.G. Edwards, S.E.J. Villar, J. Jehlicka and T. Munshi, "FT–Raman spectroscopic study of calcium-rich and magnesium-rich carbonate minerals," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 61**(10)**, pp. 2273-2280, (2005).

58. G. Vergote, T. De Beer, C. Vervaet, J.P. Remon, W. Baeyens, N. Diericx and F. Verpoort, "In-line monitoring of a pharmaceutical blending process using FT-raman spectroscopy," *European Journal of Pharmaceutical Sciences*, vol. 21**(4)**, pp. 479-485, (2004).

59. M.J. Deen and P.K. Basu, *Silicon Photonics: Fundamentals and Devices*, John Wiley & Sons, (2012).

60. Hamamatsu Photonics. *http://www.hamamatsu.com/jp/en/product/category/3100/3001/index.html* [Online].

61. Hamamatsu Photonics. *http://www.hamamatsu.com/resources/pdf/etd/R11102_-01_TPMH1324E02.pdf* [Online].

62. B.E. Saleh and M.C. Teich, "Fundamental of photonics," *Wiley Online Library*, (2007).

63. Q. Wu, T. Hamilton, W.H. Nelson, S. Elliott, J.F. Sperry and M. Wu, "UV raman spectral intensities of E. coli and other bacteria excited at 228.9, 244.0, and 248.2 nm," *Anal.Chem.*, vol. 73**(14)**, pp. 3432-3440, (2001).

64. T. Iliescu, M. Baia and V. Miclăuş, "A raman spectroscopic study of the diclofenac sodium–β-cyclodextrin interaction," *European journal of pharmaceutical sciences*, vol. 22**(5)**, pp. 487-495, (2004).

65. Princeton Instruments. *http://www.princetoninstruments.com/products/speccam/PyLoN/dsheet.aspx* [Online].

66. S.M. Sze and K.K. Ng, *Physics of semiconductor devices*, New York: Wiley-Interscience, (2006).

67. L. Pancheri, M. Scandiuzzo, D. Stoppa and G. Betta, "Low-noise avalanche photodiode in standard 0.35-µm CMOS technology," *IEEE Trans. Electron Devices*, vol. 55**(1)**, pp. 457-461, (2008).

68. M. Lee and W. Choi, "A silicon avalanche photodetector fabricated with standard CMOS technology with over 1 THz gain-bandwidth product," *Optics Express*, vol. 18**(23)**, pp. 24189-24194, (2010).

69. D. Cutler, "Fourier transform raman instrumentation," *Spectrochim.Acta, Pt.A: Mol.Spectrosc.*, vol. 46**(2)**, pp. 131-151, (1990).

70. I.R. Lewis and H. Edwards, *Handbook of Raman spectroscopy: from the research laboratory to the process line*, CRC Press, (2001).

71. A. Wainright, S.J. Williams, G. Ciambrone, Q. Xue, J. Wei and D. Harris, "Sample pre-concentration by isotachophoresis in microfluidic devices," *Journal of Chromatography A*, vol. 979**(1)**, pp. 69-80, (2002).

72. J.P. Lafleur, A.A. Rackov, S. McAuley and E.D. Salin, "Miniaturised centrifugal solid phase extraction platforms for in-field sampling, pre-concentration and spectrometric detection of organic pollutants in aqueous samples," *Talanta*, vol. 81**(1)**, pp. 722-726, (2010).

73. I.C. Nwaneshiudu, Q. Yu and D.T. Schwartz, "Quantitative solid-phase microextraction (SPME)-raman spectroscopy for the detection of trace organics in water," *Appl.Spectrosc.*, vol. 66**(12)**, pp. 1487-1491, (2012).

74. P.L. Stiles, J.A. Dieringer, N.C. Shah and R.P. Van Duyne, "Surface-enhanced raman spectroscopy," *Annu.Rev.Anal.Chem.*, vol. 1, pp. 601-626, (2008).

75. M. Fan, G.F. Andrade and A.G. Brolo, "A review on the fabrication of substrates for surface enhanced raman spectroscopy and their applications in analytical chemistry," *Anal.Chim.Acta*, vol. 693**(1)**, pp. 7-25, (2011).

76. R.A. Halvorson and P.J. Vikesland, "Surface-enhanced raman spectroscopy (SERS) for environmental analyses," *Environ.Sci.Technol.*, vol. 44**(20)**, pp. 7749-7755, (2010).

77. M. Fleischmann, P. Hendra and A. McQuillan, "Raman spectra of pyridine adsorbed at a silver electrode," *Chemical Physics Letters*, vol. 26**(2)**, pp. 163-166, (1974).

78. K. Kneipp, Y. Wang, H. Kneipp, L.T. Perelman, I. Itzkan, R.R. Dasari and M.S. Feld, "Single molecule detection using surface-enhanced raman scattering (SERS)," *Phys.Rev.Lett.*, vol. 78**(9)**, pp. 1667-1670, (1997).

79. S. Nie and S.R. Emory, "Probing single molecules and single nanoparticles by surface-enhanced raman scattering," *Science*, vol. 275**(5303)**, pp. 1102-1106, (1997).

80. N.A. Abu Hatab, J.M. Oran and M.J. Sepaniak, "Surface-enhanced raman spectroscopy substrates created via electron beam lithography and nanotransfer printing," *Acs Nano*, vol. 2**(2)**, pp. 377-385, (2008).

81. K. Kneipp, A.S. Haka, H. Kneipp, K. Badizadegan, N. Yoshizawa, C. Boone, K.E. Shafer-Peltier, J.T. Motz, R.R. Dasari and M.S. Feld, "Surface-enhanced raman spectroscopy in single living cells using gold nanoparticles," *Appl.Spectrosc.*, vol. 56**(2)**, pp. 150-154, (2002).

82. S. Lee, J. Choi, L. Chen, B. Park, J.B. Kyong, G.H. Seong, J. Choo, Y. Lee, K. Shin and E.K. Lee, "Fast and sensitive trace analysis of malachite green using a surface-enhanced raman microfluidic sensor," *Anal.Chim.Acta*, vol. 590**(2)**, pp. 139-144, (2007).

83. K. Yea, S. Lee, J.B. Kyong, J. Choo, E.K. Lee, S. Joo and S. Lee, "Ultra-sensitive trace analysis of cyanide water pollutant in a PDMS microfluidic channel using surface-enhanced raman spectroscopy," *Analyst*, vol. 130, pp. 1009, (2005).

84. X. Zhang, M.A. Young, O. Lyandres and R.P. Van Duyne, "Rapid detection of an anthrax biomarker by surface-enhanced raman spectroscopy," *J.Am.Chem.Soc.*, vol. 127**(12)**, pp. 4484-4489, (2005).

85. P. Mosier-Boss and S. Lieberman, "Detection of anions by normal raman spectroscopy and surface-enhanced raman spectroscopy of cationic-coated substrates," *Appl.Spectrosc.*, vol. 57**(9)**, pp. 1129-1137, (2003).

86. A. Sengupta, M. Mujacic and E.J. Davis, "Detection of bacteria by surface-enhanced raman spectroscopy," *Analytical and bioanalytical chemistry*, vol. 386**(5)**, pp. 1379-1386, (2006).

87. C. Ruan, W. Luo, W. Wang and B. Gu, "Surface-enhanced raman spectroscopy for uranium detection and analysis in environmental samples," *Anal.Chim.Acta*, vol. 605**(1)**, pp. 80-86, (2007).

88. N.A. Hatab, G. Eres, P.B. Hatzinger and B. Gu, "Detection and analysis of cyclotrimethylenetrinitramine (RDX) in environmental samples by surface-enhanced raman spectroscopy," *J.Raman Spectrosc.*, vol. 41**(10)**, pp. 1131-1136, (2010).

89. J. Yan, X. Han, J. He, L. Kang, B. Zhang, Y. Du, H. Zhao, C. Dong, H. Wang and P. Xu, "Highly sensitive surface-enhanced raman spectroscopy (SERS) platforms based on silver nanostructures fabricated on polyaniline membrane surfaces," *ACS Applied Materials & Interfaces*, vol. 4**(5)**, pp. 2752-2756, (2012).

90. L. He, N. Kim, H. Li, Z. Hu and M. Lin, "Use of a fractal-like gold nanostructure in surface-enhanced raman spectroscopy for detection of selected food contaminants," *J.Agric.Food Chem.*, vol. 56**(21)**, pp. 9843-9847, (2008).

91. J. Hu, P. Zheng, J. Jiang, G. Shen, R. Yu and G. Liu, "Electrostatic interaction based approach to thrombin detection by surface-enhanced raman spectroscopy," *Anal.Chem.*, vol. 81**(1)**, pp. 87-93, (2008).

92. M. Mulvihill, A. Tao, K. Benjauthrit, J. Arnold and P. Yang, "Surface-Enhanced raman spectroscopy for trace arsenic detection in contaminated water," *Angewandte Chemie*, vol. 120**(34)**, pp. 6556-6560, (2008).

93. S. Tan, M. Erol, S. Sukhishvili and H. Du, "Substrates with discretely immobilized silver nanoparticles for ultrasensitive detection of anions in water using surface-enhanced raman scattering," *Langmuir*, vol. 24**(9)**, pp. 4765-4771, (2008).

94. W. Wang and B. Gu, "New surface-enhanced raman spectroscopy substrates via self-assembly of silver nanoparticles for perchlorate detection in water," *Appl.Spectrosc.*, vol. 59**(12)**, pp. 1509-1515, (2005).

95. K.C. Bantz and C.L. Haynes, "Surface-enhanced raman scattering detection and discrimination of polychlorinated biphenyls," *Vibrational Spectroscopy*, vol. 50**(1)**, pp. 29-35, (2009).

96. C. Ruan, W. Wang and B. Gu, "Surface-enhanced raman scattering for perchlorate detection using cystamine-modified gold nanoparticles," *Anal.Chim.Acta*, vol. 567**(1)**, pp. 114-120, (2006).

97. D. Bhandari, S.M. Wells, S.T. Retterer and M.J. Sepaniak, "Characterization and detection of uranyl ion sorption on silver surfaces using surface enhanced raman spectroscopy," *Anal.Chem.*, vol. 81**(19)**, pp. 8061-8067, (2009).

98. E. Ippen and C. Shank, "Picosecond response of a high− repetition− rate CS optical kerr gate," *Appl.Phys.Lett.*, vol. 26, pp. 92, (1975).

99. P. Matousek, M. Towrie, A. Stanley and A. Parker, "Efficient rejection of fluorescence from raman spectra using picosecond kerr gating," *Appl.Spectrosc.*, vol. 53**(12)**, pp. 1485-1489, (1999).

100. F. Knorr, Z.J. Smith and S. Wachsmann-Hogiu, "Development of a time-gated system for raman spectroscopy of biological samples," *Optics Express*, vol. 18**(19)**, pp. 20049-20058, (2010).

101. M.C.H. Prieto, P. Matousek, M. Towrie, A.W. Parker, M. Wright, A.W. Ritchie and N. Stone, "Use of picosecond kerr-gated raman spectroscopy to suppress signals from both surface and deep layers in bladder and prostate tissue," *J.Biomed.Opt.*, vol. 10**(4)**, pp. 044006-044006-6, (2005).

102. B. Rebecca, M. Pavel and R. Kate, "Depth profiling of calcifications in breast tissue using picosecond kerr-gated raman spectroscopy," *Analyst*, vol. 132**(1)**, pp. 48-53, (2007).

103. E.V. Efremov, J.B. Buijs, C. Gooijer and F. Ariese, "Fluorescence rejection in resonance raman spectroscopy using a picosecond-gated intensified charge-coupled device camera," *Appl.Spectrosc.*, vol. 61**(6)**, pp. 571-578, (2007).

104. Y. Fleger, L. Nagli, M. Gaft and M. Rosenbluh, "Narrow gated raman and luminescence of explosives," *J Lumin*, vol. 129**(9)**, pp. 979-983, (2009).

105. F. Ariese, H. Meuzelaar, M.M. Kerssens, J.B. Buijs and C. Gooijer, "Picosecond raman spectroscopy with a fast intensified CCD camera for depth analysis of diffusely scattering media," *Analyst*, vol. 134**(6)**, pp. 1192-1197, (2009).

106. J.V. Sinfield, O. Colic, D. Fagerman and C. Monwuba, "A low cost time-resolved raman spectroscopic sensing system enabling fluorescence rejection," *Appl.Spectrosc.*, vol. 64**(2)**, pp. 201-210, (2010).

107. J.C. Carter, S.M. Angel, M. Lawrence-Snyder, J. Scaffidi, R.E. Whipple and J.G. Reynolds, "Standoff detection of high explosive materials at 50 meters in ambient light conditions using a small raman instrument," *Appl.Spectrosc.*, vol. 59**(6)**, pp. 769-775, (2005).

108. N. Faramarzpour, M. El-Desouki, M.J. Deen, Qiyin Fang, S. Shirani and L.W.C. Liu, "CMOS imaging for biomedical applications," *IEEE Potentials*, vol. 27**(3)**, pp. 31-36, (2008).

109. N. Faramarzpour, M.J. Deen and S. Shirani, "An approach to improve the signal-to-noise ratio of active pixel sensor for low-light-level applications," *IEEE Trans. Electron Devices*, vol. 53**(9)**, pp. 2384-2391, (2006).

110. Y. Ardeshirpour, M.J. Deen and S. Shirani, "Evaluation of complementary metal-oxide semiconductor based photodetectors for low-level light detection," *Journal of Vacuum Science & Technology A*, vol. 24**(3)**, pp. 860-865, (2006).

111. M. Kfouri, O. Marinov, P. Quevedo, N. Faramarzpour, S. Shirani, L. Liu, Q. Fang and M.J. Deen, "Toward a miniaturized wireless fluorescence-based diagnostic imaging system," *IEEE J. Sel. Topics Quantum Electron.*, vol. 14**(1)**, pp. 226-234, (2008).

112. A.P. Esposito, C.E. Talley, T. Huser, C.W. Hollars, C.M. Schaldach and S.M. Lane, "Analysis of single bacterial spores by micro-raman spectroscopy," *Appl.Spectrosc.*, vol. 57**(7)**, pp. 868-871, (2003).

113. M.F. Escoriza, J.M. VanBriesen, S. Stewart, J. Maier and P.J. Treado, "Raman spectroscopy and chemical imaging for quantification of filtered waterborne bacteria," *J.Microbiol.Methods*, vol. 66**(1)**, pp. 63-72, (2006).

114. C. Camerlingo, I. Delfino, G. Perna, V. Capozzi and M. Lepore, "Micro-raman spectroscopy and univariate analysis for monitoring disease follow-up," *Sensors*, vol. 11**(9)**, pp. 8309-8322, (2011).

115. E.G. Loewen and E. Popov, *Diffraction gratings and applications*, CRC Press, (1997).

116. K.K. Sharma, *Optics: principles and applications*, Academic Press, (2006).

117. PCGrate. *http://www.pcgrate.com/about/company* [Online].

118. A.C. Thompson and Vaudhn D, *X-Ray Data Booklet*, 2nd ed., Lawrence Berkeley Laboratory, (2001).

119. E. Loewen G, *Diffraction grating handbook*, 1 ed. USA, Newport Corporation, (2005).

120. M.M. El-Desouki, D. Palubiak, M. Deen, Q. Fang and O. Marinov, "A novel, high-dynamic-range, high-speed, and high-sensitivity CMOS imager using time-domain single-photon counting and avalanche photodiodes," *IEEE Sens. J.*, vol. 11**(4)**, pp. 1078-1083, (2011).

121. M.M. El-Desouki, O. Marinov, M. Deen and Q. Fang, "CMOS active-pixel sensor with in-situ memory for ultrahigh-speed imaging," *IEEE Sens. J.*, vol. 11**(6)**, pp. 1375-1379, (2011).

122. D.P. Palubiak and M.J. Deen, "CMOS SPADs: Design issues and research challenges for detectors, circuits, and arrays," *IEEE J. Sel. Topics Quantum Electron.*, (2014).

123. D. Li, J. Arlt, J. Richardson, R. Walker, A. Buts, D. Stoppa, E. Charbon and R. Henderson, "Real-time fluorescence lifetime imaging system with a 32× 32 0.13 μm CMOS low dark-count single-photon avalanche diode array," *Optics Express*, vol. 18**(10)**, pp. 10257-10269, (2010).

124. A. Dalla Mora, A. Tosi, F. Zappa, S. Cova, D. Contini, A. Pifferi, L. Spinelli, A. Torricelli and R. Cubeddu, "Fast-gated single-photon avalanche diode for wide dynamic range near infrared spectroscopy," *IEEE J. Sel. Topics Quantum Electron.*, vol. 16**(4)**, pp. 1023-1030, (2010).

125. C. Niclass, M. Soga, H. Matsubara, M. Ogawa and M. Kagami, "A 0.18μm CMOS SoC for a 100m-range 10fps 200× 96-pixel time-of-flight depth sensor," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2013 IEEE International*, Jan. 2013, pp. 488-489.

126. Y. Maruyama, J. Blacksberg and E. Charbon, "A 1024 x 8, 700-ps time-gated SPAD line sensor for planetary surface exploration with laser raman spectroscopy and LIBS," *IEEE J. Solid-State Circuits*, vol. 49, pp. 179-189, (2014).

127. J. Kostamovaara, J. Tenhunen, M. Kögler, I. Nissinen, J. Nissinen and P. Keränen, "Fluorescence suppression in raman spectroscopy using a time-gated CMOS SPAD," *Optics Express*, vol. 21**(25)**, pp. 31632-31645, (2013).

128. I. Nissinen, J. Nissinen, A. Lansman, L. Hallman, A. Kilpela, J. Kostamovaara, M. Kogler, M. Aikio and J. Tenhunen, "A sub-ns time-gated CMOS single photon avalanche diode detector for Raman spectroscopy," in *Solid-State Device Research Conference (ESSDERC), 2011 Proceedings of the European*, 2011, pp. 375-378.

129. Z. Li and M.J. Deen, "Towards a portable raman spectrometer using a concave grating and a time-gated CMOS SPAD," *Optics Express*, vol. 22**(15)**, pp. 18736-18747, (2014).

130. E. Charbon, "Single-photon imaging in complementary metal oxide semiconductor processes," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 372**(2012)**, pp. 20130100, (2014).

131. A. Gallivanoni, I. Rech and M. Ghioni, "Progress in quenching circuits for single photon avalanche diodes," *IEEE Trans. Nuclear Science*, vol. 57**(6)**, pp. 3815-3826, (2010).

132. A. Eisele, R. Henderson, B. Schmidtke, T. Funk, L. Grant, J. Richardson and W. Freude, "185 MHz count rate, 139 dB dynamic range single-photon avalanche diode with active quenching circuit in 130 nm CMOS technology," in *Int. Image Sensor Workshop (IISW), Onuma, Hokkaido*, 2011.

133. D. Palubiak, M.M. El-Desouki, O. Marinov, M. Deen and Q. Fang, "High-speed, single-photon avalanche-photodiode imager for biomedical applications," *IEEE Sens. J.*, vol. 11**(10)**, pp. 2401-2412, (2011).

134. A. Dalla Mora, A. Tosi, F. Zappa, S. Cova, D. Contini, A. Pifferi, L. Spinelli, A. Torricelli and R. Cubeddu, "Fast-gated single-photon avalanche diode for wide dynamic range near infrared spectroscopy," *IEEE J. Sel. Topics Quantum Electron.*, vol. 16**(4)**, pp. 1023-1030, (2010).

135. J. Sun and P. Wang, "On-chip blumlein pulse generator," *Electron.Lett.*, vol. 50**(12)**, pp. 859-860, (2014).

136. J. Sun and P. Wang, "Note: Complementary metal–oxide–semiconductor high voltage pulse generation circuits," *Rev.Sci.Instrum.*, vol. 84**(10)**, pp. 106111, (2013).

137. J. Sun and P. Wang, "A CMOS short pulse generator with a high-voltage stacked MOSFET switch," in *Circuits and Systems (MWSCAS), 2014 IEEE 57th International Midwest Symposium on*, 2014, pp. 414-417.

138. M. Ghioni, A. Gulinatti, I. Rech, F. Zappa and S. Cova, "Progress in silicon single-photon avalanche diodes," *IEEE J. Sel. Topics Quantum Electron.*, vol. 13**(4)**, pp. 852-862, (2007).

139. G. Hurkx, H. De Graaff, W. Kloosterman and M. Knuvers, "A new analytical diode model including tunneling and avalanche breakdown," *IEEE Trans. Electron Devices*, vol. 39**(9)**, pp. 2090-2098, (1992).

140. C. Ma, M. Deen and L. Tarof, "Characterization and modeling of SAGCM InP/InGaAs avalanche photodiodes for multigigabit optical fiber," *Advances in Imaging and Electron Physics*, vol. 99, pp. 65, (1997).

141. C.L.F. Ma, M.J. Deen, L.E. Tarof and J.C.H. Yu, "Temperature dependence of breakdown voltages in separate absorption, grading, charge, and multiplication InP/InGaAs avalanche photodiodes," *IEEE Trans. Electron Devices*, vol. 42**(5)**, pp. 810-818, (1995).

142. R. Pagano, D. Corso, S. Lombardo, G. Valvo, D.N. Sanfilippo, G. Fallica and S. Libertino, "Dark current in silicon photomultiplier pixels: Data and model," *IEEE Trans. Electron Devices*, vol. 59**(9)**, pp. 2410-2416, (2012).

143. R. Pagano, G. Valvo, D. Sanfilippo, S. Libertino, D. Corso, P. Fallica and S. Lombardo, "Silicon photomultiplier device architecture with dark current improved to the ultimate physical limit," *Appl.Phys.Lett.*, vol. 102**(18)**, pp. 183502-(1-4), (2013).

144. D. Stoppa, D. Mosconi, L. Pancheri and L. Gonzo, "Single-photon avalanche diode CMOS sensor for time-resolved fluorescence measurements," *IEEE Sens. J.*, vol. 9**(9)**, pp. 1084-1090, (2009).

145. N. Faramarzpour, M.J. Deen, S. Shirani and Q. Fang, "Fully integrated single photon avalanche diode detector in standard CMOS 0.18-µm technology," *IEEE Trans. Electron Devices*, vol. 55**(3)**, pp. 760-767, (2008).

146. S. Chick, R. Coath, R. Sellahewa, R. Turchetta, T. Leitner and A. Fenigstein, "Dead time compensation in CMOS single photon avalanche diodes with active quenching and external reset," *IEEE Trans. Electron Devices*, vol. 61**(8)**, pp. 2725-2731, (2014).

147. C. Niclass, M. Gersbach, R. Henderson, L. Grant and E. Charbon, "A single photon avalanche diode implemented in 130-nm CMOS technology," *IEEE J. Sel. Topics Quantum Electron.*, vol. 13**(4)**, pp. 863-869, (2007).

148. J.A. Richardson, L.A. Grant and R.K. Henderson, "Low dark count single-photon avalanche diode structure compatible with standard nanometer scale CMOS technology," *IEEE Photon. Technol. Lett.*, vol. 21**(14)**, pp. 1020-1022, (2009).

149. S. Cova, M. Ghioni, A. Lotito, I. Rech and F. Zappa, "Evolution and prospects for single-photon avalanche diodes and quenching circuits," *Journal of Modern Optics*, vol. 51**(9-10)**, pp. 1267-1288, (2004).

150. A. Eisele, R. Henderson, B. Schmidtke, T. Funk, L. Grant, J. Richardson and W. Freude, "185 MHz count rate, 139 dB dynamic range single-photon avalanche diode with active quenching circuit in 130 nm CMOS technology," in *Int. Image Sensor Workshop (IISW), Onuma, Hokkaido*, 2011, pp. 278-280.

151. D. Bronzi, S. Tisa, F. Villa, S. Bellisai, A. Tosi and F. Zappa, "Fast sensing and quenching of CMOS SPADs for minimal afterpulsing effects," *IEEE Photon. Technol. Lett.*, vol. 25**(8)**, pp. 776-779, (2013).

152. R.J. Baker, *CMOS: circuit design, layout, and simulation*, John Wiley & Sons, (2011).

153. A. Gola, L. Pancheri, C. Piemonte and D. Stoppa, "A SPAD-based hybrid system for time-gated fluorescence measurements," in *SPIE Defense, Security, and Sensing*, 2011, pp. 803315-1-14.

154. L. Pancheri and D. Stoppa, "Low-noise single photon avalanche diodes in 0.15 µm CMOS technology," in *Solid-State Device Research Conference (ESSDERC), 2011 Proceedings of the European*, 2011, pp. 179-182.

155. C. Niclass, M. Sergio and E. Charbon, "A single photon avalanche diode array fabricated in 0.35-µm CMOS and based on an event-driven readout for TCSPC experiments," in *Optics East 2006*, 2006, pp. 63720S-12.

156. E. Vilella, O. Alonso, A. Montiel, A. Vilà and A. Dieguez, "A low-noise time-gated single-photon detector in a HV-CMOS technology for triggered imaging," *Sensors and Actuators A: Physical*, vol. 201, pp. 342-351, (2013).

157. S. Tisa, F. Guerrieri and F. Zappa, "Variable-load quenching circuit for single-photon avalanche diodes," *Optics Express*, vol. 16**(3)**, pp. 2232-2244, (2008).

158. J. Zhang, R. Thew, C. Barreiro and H. Zbinden, "Practical fast gate rate InGaAs/InP single-photon avalanche photodiodes," *Appl.Phys.Lett.*, vol. 95**(9)**, pp. 091103, (2009).

159. M. Ghioni, S. Cova, F. Zappa and C. Samori, "Compact active quenching circuit for fast photon counting with avalanche photodiodes," *Rev.Sci.Instrum.*, vol. 67**(10)**, pp. 3440-3448, (1996).

160. J.F. Orgiazzi, "Packaging and characterization of NbN superconducting nanowire single photon detectors," *University of Waterloo* (2009).

161. M. Assanelli, A. Ingargiola, I. Rech, A. Gulinatti and M. Ghioni, "Photon-timing jitter dependence on injection position in single-photon avalanche diodes," *IEEE J. Quantum Electronics*, vol. 47**(2)**, pp. 151-159, (2011).

162. A. Gallivanoni, I. Rech, D. Resnati, M. Ghioni and S. Cova, "Monolithic active quenching and picosecond timing circuit suitable for large-area single-photon avalanche diodes," *Optics Express*, vol. 14**(12)**, pp. 5021-5030, (2006).

163. ISS Focus and Discover. *http://www.iss.com/resources/reference/data_tables/LifetimeDataFluorophores.html* [Online].

164. N. Boens, W. Qin, N. Basaric, J. Hofkens, M. Ameloot, J. Pouget, J. Lefevre, B. Valeur, E. Gratton and M. VandeVen, "Fluorescence lifetime standards for time and frequency domain fluorescence spectroscopy," *Anal.Chem.*, vol. 79**(5)**, pp. 2137-2149, (2007).

165. T. Vo-Dinh, L.R. Allain and D.L. Stokes, "Cancer gene detection using surface-enhanced raman scattering (SERS)," *J.Raman Spectrosc.*, vol. 33**(7)**, pp. 511-516, (2002).

166. J.W. Robinson, E.S. Frame and G.M. Frame II, *Undergraduate instrumental analysis*, CRC Press, (2014).

167. C. Niclass, C. Favi, T. Kluter, F. Monnier and E. Charbon, "Single-photon synchronous detection," *IEEE J. Solid-State Circuits*, vol. 44**(7)**, pp. 1977-1989, (2009).