

Springer Series in Optical Sciences 223

Paulo Ribeiro
Maria Raposo *Editors*

Optics, Photonics and Laser Technology 2018



Springer

Springer Series in Optical Sciences

Volume 223

Founding Editor

H. K. V. Lotsch, Nußloch, Baden-Württemberg, Germany

Editor-in-Chief

William T. Rhodes, Florida Atlantic University, Boca Raton, FL, USA

Series Editors

Ali Adibi, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Toshimitsu Asakura, Toyohira-ku, Hokkai-Gakuen University, Sapporo, Hokkaido, Japan

Theodor W. Hänsch, Max Planck Institute of Quantum Optics, Garching, Bayern, Germany

Ferenc Krausz, Max Planck Institute of Quantum Optics, Garching, Bayern, Germany

Barry R. Masters, Cambridge, MA, USA

Katsumi Midorikawa, Laser Tech Lab, RIKEN Advanced Science Institute, Saitama, Japan

Herbert Venghaus, Fraunhofer Institute for Telecommunications, Berlin, Germany

Horst Weber, Berlin, Germany

Harald Weinfurter, München, Germany

Springer Series in Optical Sciences is led by Editor-in-Chief William T. Rhodes, Florida Atlantic University, USA, and provides an expanding selection of research monographs in all major areas of optics:

- lasers and quantum optics
- ultrafast phenomena
- optical spectroscopy techniques
- optoelectronics
- information optics
- applied laser technology
- industrial applications and
- other topics of contemporary interest.

With this broad coverage of topics the series is useful to research scientists and engineers who need up-to-date reference books.

More information about this series at <http://www.springer.com/series/624>

Paulo Ribeiro · Maria Raposo
Editors

Optics, Photonics and Laser Technology 2018

 Springer

Editors

Paulo Ribeiro 
NOVA School of Science and Technology
Universidade Nova de Lisboa
Caparica, Portugal

Maria Raposo
NOVA School of Science and Technology
Universidade Nova de Lisboa
Caparica, Portugal

ISSN 0342-4111 ISSN 1556-1534 (electronic)
Springer Series in Optical Sciences
ISBN 978-3-030-30112-5 ISBN 978-3-030-30113-2 (eBook)
<https://doi.org/10.1007/978-3-030-30113-2>

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Organization

Conference Chair

Paulo Ribeiro, CEFITEC/FCT/UNL, Portugal

Program Chair

Maria Raposo, CEFITEC, FCT/UNL, Portugal

Program Committee

Luís Pereira, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal

Darine Abi Haidar, Laboratoire IMNC/Université Paris Diderot, France

Krzysztof Abramski, Laser and Fiber Electronics Group, Wrocław University of Science and Technology, Poland

Jean-Luc Adam, Université de Rennes 1—CNRS, France

Pablo Albella, University of Cantabria, Spain

Marcio Alencar, Universidade Federal de Sergipe, Brazil

Augusto Alfonso, Instituto de Ciencia y Tecnología de Materiales, Universidad de La Habana, Cuba

Tatiana Alieva, Universidad Complutense de Madrid, Spain

Gustavo Amaral, Pontifical Catholic University of Rio de Janeiro, Brazil

Jhon Avila, Universidade Federal de Sergipe, Brazil

Benfeng Bai, Tsinghua University, China

Francesco Baldini, Istituto di Fisica Applicata “Nello Carrara”, Italy

John Ballato, Clemson University, USA

István Bányász, Wigner Research Centre for Physics, Hungary
Almut Beige, University of Leeds, UK
Wojtek Bock, Université du Québec Outaouais, Canada
Antonella Bogoni, Scuola Superiore Sant’Anna-TeCIP, Italy
Sonia Boscolo, Aston University, UK
Giovanni Breglio, Università degli Studi di Napoli Federico II, Italy
Tom Brown, University of St. Andrews, UK
Luigi Bruno, Università della Calabria, Italy
Alexandre Cabral, Laboratory of Optics, Lasers and Systems (LOLS), Faculdade de Ciências da Universidade de Lisboa, Portugal
John Canning, University of Technology, Sydney, Australia
Adolfo Cartaxo, Instituto de Telecomunicações, ISCTE—Instituto Universitário de Lisboa, Portugal
Maurizio Casalino, IMM, Italian National Research Council (CNR), Italy
Eric Cassan, Centre de Nanosciences et de Nanotechnologie (UMR 9001), Université Paris-Sud, Orsay, France
Zhao Changming, Beijing Institute of Technology, China
Sima Chaotan, Huazhong University of Science and Technology, China
Jian Chen, Lumentum, USA
Xianfeng Chen, Shanghai Jiao Tong University, China
Dominique Chiaroni, Alcatel-Lucent Bell Labs, France
João Coelho, Faculdade de Ciências da Universidade de Lisboa, Portugal
Lorenzo Colace, University of Rome “Roma Tre”, Italy
Pere Colet, IFISC (Institute for Cross-Disciplinary Physics and Complex Systems), Spain
Elisabetta Collini, University of Padova, Italy
Giuseppe Coppola, Consiglio Nazionale delle Ricerche (IMM-CNR), Italy
Manuel Costa, Universidade do Minho, Portugal
Rinaldo Cubeddu, Politecnico di Milano, Italy
Andrea Cusano, University of Sannio, Italy
Judith Dawes, Macquarie University, Australia
Ignacio del Villar, Universidad Pública de Navarra, Spain
Antonio Delgado, Friedrich-Alexander University Erlangen-Nuremberg, Germany
Francesco Della Corte, “Mediterranea” University of Reggio Calabria, Italy
Ivan Divliansky, University of Central Florida—College of Optics and Photonics—CREOL, USA
John Donegan, Trinity College, Ireland
Dominik Dorosz, AGH University of Science and Technology, Poland
Manuel Dos Santos, Centro de Física e Investigação Tecnológica, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal
Arnaud Dubois, Laboratoire Charles Fabry, Institut d’Optique, France
Abdulhakem Elezzabi, University of Alberta, Canada
Sergio Fantini, Tufts University, USA
Robert Ferguson, National Physical Laboratory, UK
Milton Sergio Fernandes de Lima, Instituto de Estudos Avancados, Brazil

Maurizio Ferrari, National Research Council of Italy (CNR), Institute for Photonics and Nanotechnologies (IFN), Italy
Paulo Fiadeiro, Universidade da Beira Interior, Portugal
José Figueiredo, Universidade de Lisboa, Portugal
Luca Fiorani, Researcher at ENEA—Professor at “Lumsa”, “Roma Tre” and “Tor Vergata” Universities, Italy
Orlando Frazão, INESC Porto, Portugal
Ivana Gasulla, Polytechnic University of Valencia, Spain
Sanka Gateva, Institute of Electronics, Bulgarian Academy of Sciences, Bulgaria
Malte Gather, University of St. Andrews, UK
Marco Gianinnetto, Politecnico di Milano, Italy
Ross Gillanders, University of St. Andrews, UK
John Girkin, Department of Physics, Biophysical Sciences Institute, Durham University, UK
Guillaume Gomard, Light Technology Institute, Germany
Yandong Gong, Institute for Infocomm Research, Singapore
Norbert Grote, Fraunhofer Heinrich Hertz Institute, Germany
David Hagan, University of Central Florida, USA
Young-Geun Han, Hanyang University, Korea, Republic of
Charles Hirlimann, Institut de Physique et Chimie des Matériaux de Strasbourg, France
Werner Hofmann, Center of Nanophotonics at Technische Universität Berlin, Germany
Mark Hopkinson, University of Sheffield, UK
Iwao Hosako, National Institute of Information and Communications Technology, Japan
Silvia Huguet, IFAC—N. Carrara Institute of Applied Physics/CNR, Italy
Nicolae Hurduc, Gheorghe Asachi Technical University of Iasi, Romania
Baldemar Ibarra-Escamilla, Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico
Dieter Jaeger, University Duisburg-Essen, Germany
Jichai Jeong, Korea University, Korea, Republic of
Shibin Jiang, AdValue Photonics, USA
Phil Jones, Optical Tweezers Group, Department of Physics and Astronomy, University College London, UK
Yoshiaki Kanamori, Tohoku University, Japan
Fumihiko Kannari, Keio University, Japan
Christian Karnutsch, Karlsruhe University of Applied Sciences, Faculty of Electrical Engineering and Information Technology, Germany
Miroslaw Karpierz, Faculty of Physics, Warsaw University of Technology, Poland
Henryk Kasprzak, Wrocław University of Science and Technology, Poland
Gerd Keiser, Photonics Comm. Solutions Inc., USA
Qian Kemaο, School of Computer Engineering, Singapore
Vincent Kermene, Université de Limoges, XLIM Research Institute, France
Stefan Kirstein, Humboldt-Universität zu Berlin, Germany

Tomasz Kozacki, Warsaw University of Technology (WUT), Institute of Micromechanics and Photonics (IMiF), Institute of Microelectronics and Optoelectronics (IMiO), Poland

Cristina Kurachi, Instituto de Física de São Carlos, Universidade de São Paulo, Brazil

Jesper Lægsgaard, Technical University of Denmark, Denmark

Franco Laeri, Technische Universität Darmstadt, Fachbereich Physik, Germany

Vasudevan Lakshminarayanan, University of Waterloo, Canada

Emma Lazzeri, Scuola Superiore Sant'Anna-TeCIP, Italy

Wei Lee, National Chiao Tung University, Taiwan

Claire Lefort, XLIM Institute, France

Dawei Liang, FCT/UNL, Portugal

Odile Liboiron-Ladouceur, McGill University, Canada

João Lima, REQUIMTE-LAQV, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa, Portugal

Yu-Lung Lo, National Cheng Kung University, Taiwan

Carmen Machuca, Technische Universität München, Germany

Boris Malomed, Tel Aviv University, Israel

Sergey Mantsevich, M. V. Lomonosov Moscow State University, Russian Federation

Alexandre Marletta, Universidade Federal de Uberlândia, Brazil

Manuel Marques, FCUP, Portugal

Cleber Mendonça, IFSC/USP, Brazil

Francisco Meseguer, Research National Council of Spain, Spain

Paulo Miranda, Universidade de São Paulo, Brazil

Silvia Mittler, University of Western Ontario, Canada

Paulo Monteiro, Instituto de Telecomunicações, Portugal

Teresa Monteiro, Universidade de Aveiro, Portugal

Vasant Natarajan, Indian Institute of Science, India

Phyllis Nelson, California State Polytechnic University Pomona, USA

Armando Neves, University of Aveiro, Portugal

André Nicolet, Institut Fresnel, Aix Marseille Université, CNRS, France

Eduardo Nunes-Pereira, University of Minho, Portugal

Luca Palmieri, University of Padua, Italy

Krassimir Panajotov, Free University of Brussels, Belgium

Ufuk Parali, Ardahan University, Turkey

Matteo Paris, Università degli Studi di Milano, Italy

YongKeun Park, Department of Physics, KAIST, Korea, Republic of

Romain Peretti, Swiss Federal Institute of Technology in Zurich (ETHZ), Switzerland

Klaus Petermann, Technische Universität Berlin, Germany

Angela Piegari, ENEA, Italy

Armando Pinto, Instituto de Telecomunicações—University of Aveiro, Portugal

Ting-Chung Poon, Virginia Tech, USA

Alexander Popov, D. V. Skobeltsyn Institute of Nuclear Physics, Moscow State University, Russian Federation
Daniel Puerto, Universidad Miguel Hernandez de Elche, Spain
Jianrong Qiu, Zhejiang University, China
Sandro Rao, University “Mediterranea” of Reggio Calabria, Italy
Maria Raposo, CEFITEC, FCT/UNL, Portugal
João Rebola, Instituto de Telecomunicações, ISCTE-IUL, Portugal
Paulo Ribeiro, CEFITEC/FCT/UNL, Portugal
David Richardson, University of Southampton, UK
Mauricio Rico, Centro de Láseres de Pulsados (CLPU), Spain
Giancarlo Righini, Enrico Fermi Centre and IFAC CNR, Italy
Ana Rocha, Instituto de Telecomunicações, Aveiro, Portugal
José Rodrigues Jr., Universidade Federal de Sergipe, Brazil
Murilo Romero, University of Sao Paulo, Brazil
Luis Roso, Centro de Laseres Pulsados, CLPU, Spain
Svilen Sabchevski, Institute of Electronics, Bulgarian Academy of Sciences, Bulgaria
José Santos, Universidade do Porto, Portugal
Michael Scalora, Aviation and Missile Research Development and Engineering Center (AMRDEC), USA
David Schmol, Université de Versailles/Saint-Quentin, GEMaC CNRS (UMR 8635), France
Marc Sentis, Aix-Marseille University, CNRS, LP3 UMR 7341, France
Susana Sérgio, CEFITEC, Faculdade de Ciências e Tecnologia, Portugal
Gaurav Sharma, University of Rochester, USA
Susana Silva, INESC TEC, Portugal
Javier Solis, Laser Processing Group—LPG, Spain
Feng Song, Nankai University, China
Viorica Stancalie, National Institute for Laser, Plasma and Radiation Physics, Romania
Dimitar Stoyanov, Institute of Electronics, Bulgarian Academy of Sciences, Bulgaria
Slawomir Sujecki, University of Nottingham, UK
João Tavares, FEUP—Faculdade de Engenharia da Universidade do Porto, Portugal
António Teixeira, Instituto de Telecomunicações, Portugal
Ladislav Tichy, University of Pardubice, Czech Republic
James Trolinger, MetroLaser, USA
Cosimo Trono, Istituto di Fisica Applicata “Nello Carrara”, CNR, Italy
Valery Tuchin, Saratov State University, Russian Federation
Eszter Udvary, Optical and Microwave Telecommunication Laboratory, Budapest University of Technology and Economics (BME), Hungary
Zinan Wang, University of Electronic Science and Technology, China
James Wilkinson, Faculty of Physical and Applied Sciences, University of Southampton, UK
Lech Wosinski, KTH—Royal Institute of Technology, Sweden

Kaikai Xu, University of Electronic Science and Technology of China (UESTC),
China

Lianxiang Yang, Oakland University, Rochester, USA

Alison Yao, University of Strathclyde, UK

Kiyotoshi Yasumoto, Kyushu University, Japan

Anna Zawadzka, Institute of Physics, Faculty of Physics, Astronomy and
Informatics, Nicolaus Copernicus University, Poland

Lin Zhang, Aston University, UK

Additional Reviewer

Rossen Todorov, Institute of Optical Materials and Technologies, Bulgarian
Academy of Sciences, Bulgaria

Invited Speakers

Maurizio Ferrari, Istituto di Fotonica e Nanotecnologie (CNR-IFN), Italy

Kevin MacDonald, Optoelectronics Research Centre, University of Southampton,
UK

Bill Graham, Queens University of Belfast, UK

Pablo Albella, University of Cantabria, Spain

Preface

This book brings together review chapters on a selected range of themes in the science of light—developed from, and inspired by, papers presented at the 6th International Conference on Photonics, Optics and Laser Technology (PHOTOPTICS 2018), held in Funchal, Madeira, from 25 to 27 January 2018. These papers have been selected by the chairs from about 80 paper submissions, originating from 28 countries, using criteria-based assessment and comments from the program committee members, feedback from the session chairs, and the program chairs’ global view of all contributions included in the technical program.

This volume includes contributions on both theoretical and practical aspects, across the fields of Optics, Photonics and Laser Technology, all of which providing an advanced understanding of current research trends. The book contents, tilled up through 11 chapters, include Imaging Techniques, with contributions as multiphase flows photonics-enhanced image-detection sensing and dictionary construction methods for hyperspectral remote sensing correlation imaging; Laser Device Technology, addressing dual-wavelength tunable C-Band SOA-based fibre lasers for continuous-wave terahertz generation and polarization-modulated vertical-cavities surface-emitting lasers in quantum key distribution; Optical Sensors and Fibre Optics embracing, surface plasmon resonance miniaturized sensor systems—opportunities and challenges, high speed mode separating fibre optical detectors, centimeter-resolution long-distance optical fibre monitoring; numerical simulation of the plasma inside a glow discharge millimeter-wave detector and finally, the field of Optical Techniques Simulation and Analysis gathers the contributions, optical scatter—techniques and analysis and numerical simulation of the plasma inside a glow discharge millimeter-wave detector.

We thank all of the authors for their contributions, and the reviewers who have greatly contributed to ensure the quality of this publication.

Caparica, Portugal

Maria Raposo
Paulo Ribeiro

Contents

1	Optical Scatter—Techniques and Analysis	1
	Schubert Soares	
1.1	Introduction	1
1.2	Optical Sensing and Media Metrology	2
1.2.1	Laser Reflection and Scatter: Defects and Surfaces	3
1.2.2	Optical Sensing for Scatterometry	10
1.2.3	Laser Triangulation Sensing	13
1.3	Fabrication of Optical and Semiconductor Surfaces	15
1.3.1	Spindle: Torque Offset and Thermal Stability	15
1.3.2	Tool-Holder Stage: Positional Accuracy and Backlash	18
1.3.3	Spindle: Dynamic Axial and Radial Instability	19
1.3.4	Spindle: Error Motion Analysis for Product Formation	21
1.3.5	Cutting Tool: Impact of Error Motion	24
1.3.6	Product Analysis: Surface, Defect, Thin Film Inspection	25
1.3.7	Scatterometry: Future Applications in the Environment	29
1.4	Conclusions	30
	References	31
2	Centimeter-Resolution Long-Distance Optical Fiber Monitoring	33
	Felipe Calliari, Luis Herrera, Jean Pierre von der Weid and Gustavo Amaral	
2.1	Introduction	33
2.2	Reflectometry Strategies for Optical Fiber Monitoring	34
2.3	Fault Location Algorithms	42

- 2.4 Long-Distance Centimeter-Resolution Monitoring Mediated by Fault Location Algorithm 44
- 2.5 Monitoring Different Optical Fiber Links 46
- 2.6 Tunability, Coherence, Chromatic Dispersion and Spatial Resolution 51
- 2.7 Conclusions 53
- References 54
- 3 Characterization of High Speed Optical Detectors by Using a Mode Separating Fiber 57**
 - F. J. Achten and D. Molin
 - 3.1 Introduction 57
 - 3.2 Characterization Method 59
 - 3.3 Mode Separating Fiber 63
 - 3.4 The Detectors 67
 - 3.5 Measurement Results 68
 - 3.5.1 Reference 68
 - 3.5.2 Detectors Under Test 69
 - 3.5.3 Discussion 70
 - 3.5.4 A Closer Look 71
 - 3.6 Conclusions 73
 - References 74
- 4 Polarization Modulated Vertical-Cavity Surface-Emitting Lasers in Quantum Key Distribution 75**
 - Ágoston Schranz and Eszter Udvary
 - 4.1 Quantum Key Distribution 75
 - 4.1.1 Necessity of QKD in Cryptography 76
 - 4.1.2 QKD Protocols and Practical Implementations 77
 - 4.1.3 Examples of Simple DV-QKD Protocols 78
 - 4.2 VCSELs in QKD Systems 79
 - 4.2.1 Advantages of VCSELs over EELs in QKD 79
 - 4.2.2 Drawback: Polarization Switching 80
 - 4.3 Polarization Modulation and a Newly Proposed BB84 Transmitter Design 84
 - 4.3.1 Proposed Design for BB84 Transmitters 84
 - 4.3.2 On-Demand Polarization Switching 85
 - 4.3.3 Difficulties in the New Design 88
 - 4.4 Spectral Attacks and Protection 89
 - 4.5 Conclusion 91
 - References 91

5	Tm Based Solid-State Lasers—Toward High Power Tunability—A Review	93
	Salman Noach, Uzziel Sheintop and Rotem Nahear	
5.1	Introduction	93
5.2	Tm Laser Characterizes	94
5.2.1	Tm ³⁺ Ion Spectroscopy	94
5.2.2	Tm Host Materials Properties	96
5.3	Tunability Methods and Results	97
5.3.1	Prism Tuning	97
5.3.2	Grating Tuning	99
5.3.3	Volume Bragg Gratings Tuning	101
5.3.4	Birefringent Filters Tuning	103
5.3.5	Febry-Perot Etalon Tuning	107
5.4	Pulsed Laser Tunability	110
5.5	Tunability Tm Lasers Overview	113
5.6	Conclusions	115
	References	115
6	A Dual-Wavelength Widely Tunable C-Band SOA-Based Fiber Laser for Continuous Wave Terahertz Generation	119
	Muhammad Ummy, Abdullah Hossain, Simeon Bikorimana and Roger Dorsinville	
6.1	Introduction	119
6.2	Various Sources for Terahertz Generation	120
6.3	Optical Sources for Terahertz Generation	121
6.3.1	Pulsed Lasers Systems	121
6.3.2	Continuous Wave Optical Laser Sources	122
6.4	Resonance Absorption Based THz Sources	123
6.4.1	Photoconductive Emitters	123
6.4.2	Semiconductor Surface Emitters	123
6.4.3	Gas Ionization	124
6.4.4	Intra-molecular Charge Generation Via Photoinduction	124
6.5	Non-resonant Nonlinear Interactions Based THz Sources	125
6.5.1	Difference Frequency Generation	125
6.6	Non-resonant Optical Rectification	126
6.6.1	Cherenkov Crystal Geometry	126
6.6.2	Excitation via Tilted Optical Pump Beams	127
6.6.3	Quasi-phase Matching	128
6.7	Novel Single Fiber Optical Laser Source Utilizing Difference Frequency Generation for Terahertz Emission	128
6.7.1	Experimental Optical Source Architecture	129
6.7.2	Dual-Wavelength Amplification and Single/Dual Output Port Operation	130

6.7.3	Dual-Wavelength Optical Beam Tunability and Optical Power Stability	132
6.7.4	Single and Dual Output Port Operation Characterization	133
6.7.5	Continuous Wave THz Emission Generation and Detection	135
6.8	Conclusion	137
	References	138
7	Reliability Challenges of Nanoscale Avalanche Photodiodes for High-Speed Fiber-Optic Communications	143
	Jack Jia-Sheng Huang, Yu-Heng Jan, H. S. Chang, C. J. Ni, Emin Chou, S. K. Lee, H. S. Chen and Jin-Wei Shi	
7.1	Introduction	143
7.1.1	Datacenter Network	144
7.1.2	Wireless Network	146
7.1.3	Passive Optical Network	147
7.1.4	Telecommunication Network	147
7.2	APD Light Detectors	148
7.2.1	Reverse IV and Dark Current	150
7.2.2	Electric Field Control	151
7.2.3	Temperature Stability	153
7.2.4	Bandwidth	154
7.2.5	Sensitivity	154
7.3	Device Miniaturization	155
7.4	Future Reliability Challenges	156
7.4.1	Optical and Electrical Overload Stress	158
7.4.2	Reliability Aging	159
7.4.3	Electrostatic Discharge	161
7.5	Conclusions	163
	References	163
8	Miniaturized Surface Plasmon Resonance Based Sensor Systems—Opportunities and Challenges	169
	Peter Hausler, Carina Roth, Thomas Vitzthumecker and Rudolf Bierl	
8.1	SPR Principle, Current Applications and Future Evolution	169
8.2	Current Approaches to Miniaturized SPR Systems	173
8.3	Technical Aspects for the Design of Miniaturized SPR Sensors	178
8.4	Outlook	192
	References	193

9	Photonics-Enhanced Image-Detection Sensing of Multiphase Flows	197
	Sergio L. Carrasco-Ortiz, Eduardo Valero, Maria Morant and Roberto Llorente	
9.1	Introduction	197
9.2	Laser-CMOS Sensor Architecture	200
	9.2.1 Photonic Subsystems	200
	9.2.2 Data Acquisition and Digital Signal Processing	203
9.3	Experimental Bubble Characterization Results	211
	9.3.1 Air Bubble Concentration Calculation	211
	9.3.2 Analysis of Bubbles' Shape	212
	9.3.3 Bubbles' Speed Calculation	213
9.4	Conclusion	217
	References	218
10	Numerical Simulation of the Plasma Inside a Glow Discharge Millimeter Wave Detector	221
	Cemre Kusoglu-Sarikaya, Demiral Akbar and Hakan Altan	
10.1	Introduction	222
10.2	Gas Discharge Plasmas	223
10.3	THz and MM-Waves	224
10.4	Glow Discharge Detectors	225
10.5	Validation of the PIC/MCC Simulation Code	228
10.6	Parallel 1d3v PIC/MCC Simulation of GDD	230
10.7	Simulation Results (Pure Neon Gas)	231
10.8	Simulation Results (Ne-Ar Mixture)	234
10.9	Conclusions	235
	References	236
11	Dictionary Construction Method for Hyperspectral Remote Sensing Correlation Imaging	239
	Qi Wang, Lingling Ma, Hong Xu, Yongsheng Zhou, Chuanrong Li, Lingli Tang and Xinhong Wang	
11.1	Introduction	240
	11.1.1 Hyperspectral Correlation Imaging	240
	11.1.2 Correlation Reconstruction via Sparse Dictionary Constraint	241
11.2	Sparse Representation and Dictionary Construction Theory	244
	11.2.1 Sparse Representation Theory	244
	11.2.2 Sparse Coding Algorithms	246
	11.2.3 Sparse Dictionary Construction Theory	249
11.3	Hyperspectral Sparse Dictionary Construction Method for Correlation Imaging	251

- 11.3.1 Sparse Representation in Hyperspectral Correlation Imaging 251
- 11.3.2 The Fixed and Sample Hyperspectral Sparse Dictionary Construction 253
- 11.3.3 Training Dictionary Construction 256
- 11.3.4 Hyperspectral Sparse Dictionary Construction of Different Class 261
- 11.4 Sparse Dictionary Selection and Optimization Method in the Reconstruction 265
 - 11.4.1 Dictionary Adaptive Selection Method in Correlation Reconstruction 265
 - 11.4.2 Optimized Adaptive Dictionary Selection Strategy Based on Spatially Neighboring Pixels 271
- 11.5 Application of Hyper-spectral Sparse Dictionary in Real Imaging System 273
 - 11.5.1 Sparse Dictionary Application in the Hyperspectral Correlation Imaging System Based on LCTF 274
 - 11.5.2 Sparse Dictionary Application in the Single Exposure Correlation Hyperspectral Imaging System 278
- 11.6 Conclusions 281
- References 283

- Author Index.** 285
- Subject Index.** 287

Contributors

F. J. Achten Prysmian Group, Eindhoven, CA, The Netherlands

Demiral Akbar Department of Mechanical Engineering, Ostim Technical University, Ankara, Turkey

Hakan Altan Middle East Technical University, Department of Physics, Ankara, Turkey

Gustavo Amaral PUC-Rio, Rio de Janeiro, Brazil

Rudolf Bierl Sensorik-ApplikationsZentrum, Ostbayerische Technische Hochschule Regensburg, Regensburg, Germany

Simeon Bikorimana The City College of New York, New York, USA

Felipe Calliari PUC-Rio, Rio de Janeiro, Brazil

Sergio L. Carrasco-Ortiz Nanophotonics Technology Center, Universitat Politècnica de València, Valencia, Spain

H. S. Chang Source Photonics, Hsinchu, Taiwan

H. S. Chen Source Photonics, Hsinchu, Taiwan

Emin Chou Source Photonics, Hsinchu, Taiwan

Roger Dorsinville The City College of New York, New York, USA

Peter Hausler Sensorik-ApplikationsZentrum, Ostbayerische Technische Hochschule Regensburg, Regensburg, Germany

Luis Herrera PUC-Rio, Rio de Janeiro, Brazil

Abdullah Hossain The City College of New York, New York, USA

Jack Jia-Sheng Huang Source Photonics, West Hills, CA, USA;
Source Photonics, Hsinchu, Taiwan

Yu-Heng Jan Source Photonics, West Hills, CA, USA;
Source Photonics, Hsinchu, Taiwan

Cemre Kusoglu-Sarikaya Middle East Technical University, Department of
Physics, Ankara, Turkey

S. K. Lee Source Photonics, Hsinchu, Taiwan

Chuanrong Li Key Laboratory of Quantitative Remote Sensing Information
Technology, Academy of Opto-Electronics, Chinese Academy of Sciences, Beijing,
China

Roberto Llorente Nanophotonics Technology Center, Universitat Politècnica de
València, Valencia, Spain

Lingling Ma Key Laboratory of Quantitative Remote Sensing Information
Technology, Academy of Opto-Electronics, Chinese Academy of Sciences, Beijing,
China

D. Molin Prysmian Group, Haisnes, France

Maria Morant Nanophotonics Technology Center, Universitat Politècnica de
València, Valencia, Spain

Rotem Nahear Department of Applied Physics, Electro-Optics Engineering
Faculty, Jerusalem College of Technology, Jerusalem, Israel

C. J. Ni Source Photonics, Hsinchu, Taiwan

Salman Noach Department of Applied Physics, Electro-Optics Engineering
Faculty, Jerusalem College of Technology, Jerusalem, Israel

Carina Roth Sensorik-ApplikationsZentrum, Ostbayerische Technische
Hochschule Regensburg, Regensburg, Germany

Ágoston Schranz Budapest University of Technology and Economics, Budapest,
Hungary

Uzziel Sheintop Department of Applied Physics, Electro-Optics Engineering
Faculty, Jerusalem College of Technology, Jerusalem, Israel

Jin-Wei Shi Department of Electrical Engineering, National Central University,
Zhongli, Taiwan

Schubert Soares Ultrafast Sensors, Brighton, CO, USA

Lingli Tang Key Laboratory of Quantitative Remote Sensing Information
Technology, Academy of Opto-Electronics, Chinese Academy of Sciences, Beijing,
China

Eszter Udvary Budapest University of Technology and Economics, Budapest,
Hungary

Muhammad Ummay New York City College of Technology, Brooklyn, USA

Eduardo Valero Nanophotonics Technology Center, Universitat Politècnica de València, Valencia, Spain

Thomas Vitzthumecker Sensorik-ApplikationsZentrum, Ostbayerische Technische Hochschule Regensburg, Regensburg, Germany

Qi Wang Key Laboratory of Quantitative Remote Sensing Information Technology, Academy of Opto-Electronics, Chinese Academy of Sciences, Beijing, China

Xinhong Wang Key Laboratory of Quantitative Remote Sensing Information Technology, Academy of Opto-Electronics, Chinese Academy of Sciences, Beijing, China

Jean Pierre von der Weid PUC-Rio, Rio de Janeiro, Brazil

Hong Xu Key Laboratory of Quantitative Remote Sensing Information Technology, Academy of Opto-Electronics, Chinese Academy of Sciences, Beijing, China

Yongsheng Zhou Key Laboratory of Quantitative Remote Sensing Information Technology, Academy of Opto-Electronics, Chinese Academy of Sciences, Beijing, China

Chapter 1

Optical Scatter—Techniques and Analysis



Schubert Soares

Abstract Optical surfaces of components in lasers, telescopes, cameras, and eye-wear, and Si wafers with nanometre devices in the semiconductor industry are fabricated to increasingly rigorous standards to improve performance, quality, and yield. This trend necessitates the development of novel techniques, innovative technology, and integrable systems for the detection and evaluation of defects and inaccuracies at nanometre levels. Surface metrology must be achieved at high-speed, utilizing synergism of rapid sampling, low noise data acquisition, analytical software, and graphics. Intelligent monitoring of various parameters, operating concurrently with fabrication processes in feedback loops, could converge industrial systems to true nanoscale production. Optical scatter is a viable inspection technique utilized in photonics and semiconductor manufacturing, where optics and Si wafers are scanned in entirety, within a matter of a few minutes, to produce detailed maps of defects, device irregularities, surface roughness, shape inaccuracies, and thin film composition. We explore advanced concepts and technology based on optical scatter measured in triangulation off a target surface with a passive sensor, integrated with nanometre scanning capability, and real-time software analysis to measure and resolve various operating parameters in industrial environments, in conjunction with the physical quality of the finished product. Our results demonstrate the limitations of contemporary manufacturing with a view to future development and process optimization, leading consistently to true nanoscale accuracy. The techniques presented herewith are tractable, and can easily be configured for the manufacture of sub- μm semiconductor devices to large meter-size optics.

1.1 Introduction

Optical products incorporating lenses, crystals, and thin films, and semiconductor wafers with active devices are manufactured to accuracies of a few nanometres. The

S. Soares (✉)

Ultrafast Sensors, 6774 E 123rd Ave, Brighton, CO 80602, USA
e-mail: soares2018@ultrafastsensors.com

© Springer Nature Switzerland AG 2019

P. Ribeiro and M. Raposo (eds.), *Optics, Photonics and Laser Technology 2018*, Springer Series in Optical Sciences 223,
https://doi.org/10.1007/978-3-030-30113-2_1

production of these components is implemented within a set framework of procedures, while metrology is carried out independently to verify the product and its manufacturing processes. Industries offering diverse commercial products require versatile, non-destructive, and integrable sensor technology to monitor fabrication in real-time, evaluate machine tool operation, and to examine final quality. The measurements encompass several tool logistics such as rotational (spindle) and linear and arcuate (stage) error motion, and cutter edge evolution and erosion during usage. Other important parameters include product surface quality and shape conformity, defects, film-layer uniformity, and manufacturing repeatability. The possibility of coordinated sensor measurements, concurrent with computer aided parametric control of the manufacturing process could result in repeatable nanoscale quality products and improved yield.

Non-destructive (contact-free) measurement techniques including capacitive and inductive sensors utilize large probe dimensions, and rely on large-area sampling, rendering them unfeasible for lateral nanometre (contrast) sensitivity, simple calibration, convenient integration, and cost benefit for a multiplicity of materials, shapes, and product sizes. Charge and current sensitive metrology systems were developed originally for conductive materials, implying that non-invasive evaluation of optics or doped semiconductors may pose significant challenges. These systems are refined for fixed geometries, presenting additional restrictions, with calibration virtually impossible for aspheric or free form surfaces, and applicability limited largely to planar products. The measurement of the contour of an edge, thin-film structure, or surface and defect analysis, for example, is quite difficult utilizing these techniques. These sensors also require amplification and signal processing, and the associated electronic noise and linearity are major limiting factors to bandwidth, accuracy, and resolution.

Optical sensing in conjunction with laser triangulation could constitute a simple yet feasible solution with a microscopic footprint, for micro-to-macro scale fabrication and products. This technique could track operating parameters in the manufacturing environment, concurrently monitor, evaluate, and analyse fabrication processes, and simultaneously measure product formation from a variety of materials. Optical-lever metrology, where the specular reflection of a laser beam and surface scatter are scanned on high-resolution solid state sensors, is presented herewith. The technology is highly versatile for multi-parameter sensing, utilizes simple calibration processes, and is compatible with standard data-acquisition and computer systems.

1.2 Optical Sensing and Media Metrology

Metrology in the optical domain utilizes the physical properties of radiation (usually from a laser source) such as its wavelength, coherence, and polarization, in electromagnetic interaction with surfaces, layers, or general media composed of a slew of materials. Meaningful data is obtained by measuring the manipulation of the physical characteristics of optical reflection, diffraction and scatter, or transmission

produced by the target. Optical probe signals are typically monitored with sensors and/or cameras, logged on data-acquisition systems, and analysed with software to produce calibrated or referenced measurements to evaluate processes, to enforce product quality, and to maintain high yield in the manufacturing environment.

The primary optical sensing technique utilized widely in several industries is interferometry: two coherent polarized beams, originating from the same fixed-wavelength laser source, split by wavefront or amplitude division, constitute reference and probe beams. The probe fraction interacts with a target (surface, layer, device, or material), is modified, and is recombined with the reference beam to produce fringe patterns, which are then measured individually, in sequence, or simultaneously and utilized to decode a variety of media parameters. The mixed signal for example could be utilized to provide metrics for displacement, three-dimensional (3D) shape, spectral content, optical properties, and the internal structure of optical media, including biological material.

The reflection and transmission of coherent polarized light incident on near-perfect surfaces, interfaces, and film layers provides the basis for metrology in ellipsometry, Brewsters angle (dielectrics), principal angle (metals), and critical angle with total internal reflection (dielectric interfaces) [1, 2]. A comparative analysis of polarization states, amplitude evanescence, and phase shifts, estimates the real and complex optical properties of media, examines sub-surface structure, or detects anomalous behaviour at interfaces.

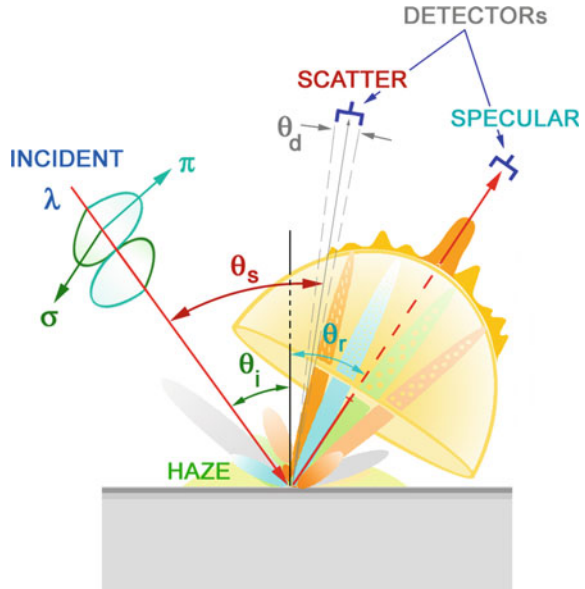
Laser-optical triangulation and surface scatter measurements, obtained by nanometre scanning, constitute novel techniques to directly image surface topography, sub-surface structure, anisotropies, artefacts, defects, and thin films, and complement static probe sensitivity to position, displacement, and motion.

1.2.1 Laser Reflection and Scatter: Defects and Surfaces

The focus of this paper is the measurement and the analysis of laser reflection and scatter from smooth surfaces in optical components and Si wafers. The physical optics of laser reflection and transmission at a perfectly smooth (defect-free and roughness-free) surface is very well understood and relies primarily on the continuity of electromagnetic fields at the interface. The important elements of the dipole and multipole natures of the lattice structure at the atomic level, in interaction with the electromagnetic field, is lumped into the permittivities, permeabilities, and the conductivities of the media but does come into active play in the development of scatter, particularly in the realm of sub-wavelength structure and defects.

A plane-polarized laser beam incident on a perfect smooth surface (Fig. 1.1) will produce specular reflection whose orientation, polarization, phase, and amplitude are governed by the scalar Snell equations and the vector Fresnel relations, which are input-polarization specific. Perturbations to this smooth surface can be introduced across the sampling beam as a single Heaviside (step) discontinuity, a slit, a small sphere, an aperture, a periodic grating, and any multi-dimensional geometric

Fig. 1.1 2π optical scatterometer: the specular beam produced by the smooth surface, is determined by the Fresnel equations. Low spatial frequency structure scatters at shallow angles, adjacent to specular reflection. High spatial frequency structure scatters away from the specular beam. The omnidirectional low-intensity haze is produced by nanostructure



combination of these elements, with superimposed distributions of residual surface roughness. The dimensions of each perturbation (and roughness) are scaled to the probe wavelength λ , and lie within the power profile in the spread of the incident beam.

The surface roughness can be ergodic, isotropic, anisotropic (with preferred orientation), fractal, etc., containing components that are sinusoidal, triangular, or random. It is generally accepted that all smooth (polished) optical and semiconductor surfaces consist of roughness error, which is considerably less than the wavelength (i.e. Rayleigh criterion ~ 0), inducing relatively insignificant scatter phase shifts. Each of the above perturbations will draw electromagnetic energy away from the incident beam and redistribute this energy apart from the specular beam, thereby changing its profile and its power structure. An associated scatter beam is generated in addition to the modified spectral beam, whose scatter power distribution, orientation, and polarization are directly correlated to the surface and defect geometry within the sampling area.

The incident beam in most systems is usually made diffuse to eliminate its coherence and minimize speckle. The ratio of specular power to scatter power is proportional to the geometry of the scattering perturbation. It is worthwhile noting here that the scatter power and distribution ranges from a mirror image specular profile (with no scatter) described by the Fresnel equations for a perfectly smooth reflecting surface comprised of identical scatterers whose structure factor vanishes, to a fully diffuse Lambertian distribution (no well defined specular beam) for a completely rough (e.g. etch) surface on the dimensional order of λ . A parallel case for uniform defect distributions also holds where large (dipole, multipole, and Mie) defects will

scatter significantly in offset orientations in comparison with near-smooth surface dipole-array scatter at sub-nanometre dimensions, making the former observable in specific scattering planes. In a fundamental unit cubic dipole array (e.g. a crystalline lattice), the internal fields on a dipole element induced by adjacent scatterers vanishes in total due to symmetry but elemental scatter at an absolute minimal level could be produced from random thermal vibrations of the lattice or by impurities (e.g. dopants in a semiconductor), and surface states.

Within the optics and the semiconductor industries, fabrication areas are highly filtered, low particle count, clean rooms with minimal human intervention. Defects and their geometries in these industries are well understood and modelled as quasi-spheres whose dimension is fractional wavelength. A defect on a semiconductor wafer is additionally classified as dielectric or metal in critical relation to its location on a bare wafer, a device pattern, or a thin film.

The topography of a rough surface can be partitioned into a smooth (zero-order) component, which represents the average of the surface function, independent of distributed roughness. The random roughness component is then generally modelled in a 3D analysis (e.g. by Fourier integral transforms) to comprehend scatter and haze. This roughness function is decomposed into a set of normalized basis functions (e.g. sinusoids) whose periodicity lies within a continuous or distributed (discrete) spread of spatial frequency components, having varying amplitudes and phase shifts, which together approximate the surface function in mean-square convergence [3–5].

The defect size, ρ , and the roughness dimension, ρ' , in dimensional comparison to the probe wavelength, λ determine the corresponding total 3D scatter distribution in Fig. 1.1 [5–7]. There are five zones of scaling, namely: $\rho' \ll \lambda$, $\rho' < \lambda$, $\rho' \sim \lambda$, $\rho' > \lambda$, and $\rho' \gg \lambda$. These zones are restricted to select categories for precision surfaces typically found in optics and semiconductor fabrication. Ar₂* excimer lasers utilized in the semiconductor industry, operating at 126 nm, could be utilized to detect dipole (defect) scatter down to the sub-10-nm device metric. At this scaling of ρ/λ , a dielectric defect on a dielectric thin film would scatter symmetrically in the 2π -azimuthal range, while a metal defect in a similar position would produce significant backscatter and minimal (~10%) forward scatter. The altitudinal scatter power minima for dielectric and metal defects occur at $\pi/2$ and $\pi/3$, respectively, for a π -polarized input beam. The positioning of detectors and the polarization of the input beam are self evident from these observations.

Referring to the scatterometer system in Fig. 1.1, with restriction to the radiation (Fraunhofer) zone in the plane of incidence, scattered optical power from a rough surface, impinges on a detector of angular spread θ_d , positioned at angular altitude θ_s , to an input beam at incidence angle θ_i . The maximum value occurs at the specular angle θ_r ($= \theta_i$), beyond which the power decays gradually in 2π -steradians if measured on a hemisphere centred at the point of reflection whose primary axis coincides with the reflected beam. In practical terms, θ_i is usually not normal, making the incident spot on the surface elliptical, which must be taken into consideration if illuminated structure or defects are large in comparison to the spot dimensions. The scatter power measurement for oblique beam incidence is compensated by a factor of $(\text{Cos } \theta_s)^{-1}$ to normalize the orientation of the detector to the plane of the surface, and the input

beam profile. The altitudinal and the azimuthal distribution of the scattered optical power around the specular axis may include higher-order maxima, dipole and Mie defect scatter, multi-tiered haze, and other stray signals.

The scaling ratio, referred to on occasion as the Rayleigh criterion, ρ'/λ determines if the optical path of scatter radiation is by direct reflection from the surface or by instances or orders of multiple reflections within the roughness, prior to exiting the surface and impinging on the detector. Low spatial-frequency roughness (several multiples of the optical wavelength; $\rho' > \lambda$) produces scatter in angular proximity to the specular beam. Nanostructure containing higher spatial frequencies ($\rho' < \lambda$) scatters power at wider angles away from the specular beam. The power of scattered radiation, in the range $\rho' \leq \lambda$, is directly proportional to the sixth power of the roughness and defect dimensions, and inversely proportional to λ^4 (Rayleigh factor). The non-specular radiation measured at oblique angles θ_s , researched extensively by Church et al. [3] and Stover [5] is Rayleigh scatter produced by densely distributed sub-wavelength $\rho' \ll \lambda$ random surface and sub-surface structure, and a high density of sub-wavelength defects.

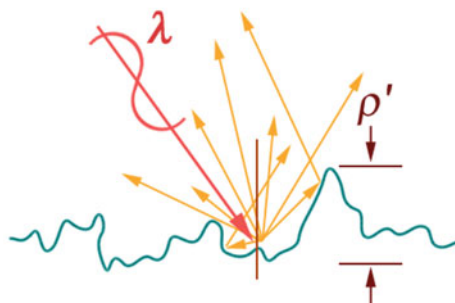
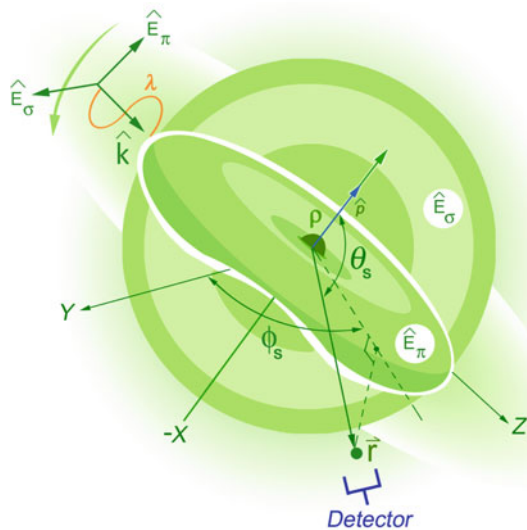
The scatter haze produced by nanometre structure (e.g. a highly polished optic or Si wafer surface), where ρ' is exceedingly insignificant in comparison with λ , is uniformly distributed over the scattering hemisphere, is visible as a Lambertian source, and forms the background noise threshold. The constituents producing this haze are irregular nanometre agglomerations at the surface with non-vanishing structure factor (e.g. amorphous structure, crystallites, or point defects), the nanometre profile of the polished surface, the nanometre sub-surface damage layer structure, multi-nanometre thin-film sputter fractal texture, low-density solid and liquid contaminants, and rarer environmental gases. Each of these sub-wavelength components produces corresponding degrees of haze, and the successful detection of a nanometre-size defect relies on the design of a system where the defect-to-haze signal ratio is greater than unity. The total power distribution function may also contain Mie scatter elements from fractional to multi-wavelength dimension structure and defects ($\rho' \sim \lambda$), and additional Rayleigh scatter from transitional range sub-wavelength defects ($\rho < \lambda$). The Rayleigh criterion may in some instances be significant (e.g. sputter fractals), resulting in small phase shifts between the incident and the scatter beams, producing a combination of speckle and high power haze. Processed semiconductor wafers possess a wide range of additional surface features such as etch texture, layer erosion and dishing (artefacts of the etch-polish process), crosshatch device patterns, step discontinuities (pads), via apertures, and debris. The rapid detection of a nanoscale defect in this environment is complex and poses very serious challenges in terms of hardware and software.

The remaining critical variables in roughness, defect, and artefact detection are the input beam angle of incidence, the degree of coherence, the polarization, the scattering cross section of the defect-base (surface or device) assembly, the beam spot power profile and size, and the detection arrangement. Successful detection of an isolated defect or stray structure is largely an empirical art, based critically on experience, a studied analysis of the defect and background combinations, astute selection of input beam parameters, and the arrangement of detectors in multiple

scattering planes, combined with target-specific software algorithms to enhance high defect-to-background contrast.

As an example, referring to Fig. 1.2, consider a free spherical dielectric defect ($\rho < \lambda$) illuminated with a plane-polarized optical beam. π -polarized light illuminating the defect generates an azimuthally (ϕ_s) symmetric lobular altitudinal dipole radiation pattern oriented orthogonal to the field and the induced dipole moment. σ -polarized light illuminating the same defect produces a uniform altitudinal (θ_s) scatter distribution relative to the scattering plane and the detector position, and is a 90° rotation of the π -field. These considerations can be implemented in hardware and analytical software to enhance the defect scatter signal if it resides on a bare or a patterned semiconductor wafer. In this respect, a π -polarized input beam in near-normal incidence, combined with azimuthally oriented and obliquely positioned detectors, performing as a spatial filter, could be utilized to probe deeply embedded defects or voids. Likewise, oblique incidence σ -polarized light combined with wide normal

Fig. 1.2 The scatter power distributions for π - and σ -polarized light illuminating a small dielectric sphere, $\rho < \lambda$. The detector positioning, θ_s , identifies the scattering plane, and the power measurement is made in an altitudinal scan from the +z-axis for both polarizations. The scatter from a rough surface is shown below, where the distribution is developed in multiple orders, and the peak Rayleigh criterion is identified as the ratio ρ'/λ

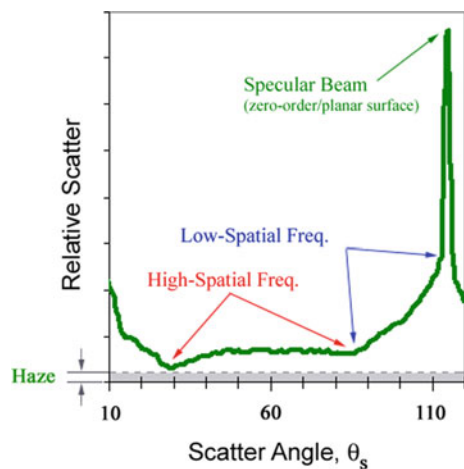


detection could be utilized to efficiently detect a small particle lying on a wafer surface. This particular arrangement also lends itself easily to segmented detection, enabling auto covariance processing or self-correlation of measured data to average out roughness scatter and improve the defect-signal to haze-background contrast.

Additional variables for consideration are the specific defect shape, which in practise is seldom spherical, its material properties, its position on a dielectric or a conducting surface, and the corresponding image charge distributions, particularly if the defect is a metal particle on a semiconductor surface or lies between conducting tracks. The operating wavelength, λ can be selected with some specificity if the background surface, the defect material, and the scatter produced in the assembly are reasonably well defined, as is generally the case in the optics and the semiconductor industries. Since scatter is distributed non-uniformly in the altitudinal and azimuthal coordinates in specific patterns, Fourier spatial filtering is predominantly engaged in the semiconductor industry to distinguish scatter power corresponding to a selected roughness spread, a particular device pattern, and a metal or dielectric defect, on processed wafers.

The typical altitudinal variation of plane-of-incidence restricted optical scatter obtained at 633 nm and π -polarization, on a highly polished quartz optic is shown in Fig. 1.3. The logarithm of the relative scatter is plotted versus the altitudinal angle in the scattering plane, which is varied from near the incident beam across to the specular reflected beam. The data follows the trends of our discussions in the previous text, and the scatter power in relation to the spread of the surface spatial frequencies is as indicated, together with the quiescent level of uniform background haze. The dynamic range is approximately 73 dB, spanning from the specular maximum power (planar surface) to the scatter minimum power (nanoroughness haze). The specific scatter power for a narrow spatial frequency spread at a prescribed angle could be resolved by precisely machining the corresponding structure onto a surface to enable a calibration reference.

Fig. 1.3 The optical scatter distribution produced by polished quartz, scanned from the incident to the specular beams at 633 nm with π -polarization. The variation of the scatter signal in relation to the spatial frequencies (scaled to the wavelength) is identified, and the dynamic range from specular power to uniform surface haze is 73 dB



A smooth-planar, polished, 100% reflecting surface produces a scatter power profile consisting primarily of the specular beam, which would have an almost identical profile to the incident beam. The noise threshold is established by a substantially minor portion of the input power transformed into Lambertian haze, produced by Rayleigh scatter at the nanoroughness level. A polished Si wafer blank, a precision single point diamond turned surface, or a highly polished optical flat or lens, are considered near-ideal surfaces, for which the surface roughness is ergodic and isotropic, and sampling the scatter produced at a single point is representative of the global surface. This can be evaluated by simply rotating the test surface around the inspection point of the input beam, or moving the beam to a second location, and measuring the stability of the scatter distribution.

We have heretofore explored scatterometry by the altitudinal resolution of the diffracted power in the scattering plane within a differential solid angle delimited by the area of the detector at its angular position. The total integrated scatter or TIS offers an alternate perspective of a surface, for which it is utilized as an associated figure of merit for comprehensive quality. The TIS forms a metric of the isotropic ergodic surface function, comprising all roughness, sub-surface damage, and defect-array constituents. The data is normalized to the specular beam (corresponding to the zero-order flat surface or roughness average), which is de-embedded from the measurement. The TIS value could be useful to monitor and to provide rapid measurements of quality, during surface preparation steps of grinding, lapping, and polishing in the process of producing a smooth flat surface on an optic or a Si wafer. It can also be configured to sample defect density per unit area by common mode comparison of the TIS value for a reference surface to a surface with distributed defects. A TIS system can be designed to scan the entire surface and provide areal metrics for a more exact monitor of overall surface preparation.

Data pertaining to surface roughness and defect distribution can be established by profiling the power in specular-scatter radiation in relation to the power profile of the incident beam to deduce the spatial frequency distribution across the surface. This technique is an affiliate to the TIS method and affords secondary means of rapid surface characterization.

Optical components, semiconductor wafers, and various other devices fabricated in precision industries are subject to in-process metrology *vis-à-vis* the analysis of parameters such as error motion, design-to-product conformity, surface topography and roughness, sub-surface damage, thickness, defects and contamination. A scatterometer can be designed to effectively limit or eliminate many of the intricacies in a 2π configuration (Fig. 1.1) contributed by factors such as variable angle of incidence, beam polarization and profile, interaction with the material, coherence, multi-channel data acquisition and processing, etc. For example, a simple scatterometer system consisting of a low-power polarized laser beam, having a symmetric mode profile, mated with passive uniaxial sensors oriented at specific angles can be utilized as a linear plane-of-incidence non-invasive probe to produce equivalent results with the advantages of rapidity, clarity, calibration ease, and austerity.

A sensor capable of resolving minute power variations proportionally enhances the corresponding scatter-probe sensitivity to surface topography, various distributed

features, and defects. As the probe is scanned across a surface, its low spatial frequency variations in topography stimulate well-defined deflections in the specular beam, and produce proportional signal fluctuations on the sensor. We have observed in some instances that the specular beam data can also be processed to identify sub-surface damage. If additional sensors are positioned at specific altitudinal orientations, within the plane of incidence, to measure scatter at predetermined non-specular angles θ_s , the detection and resolution of various surface parameters such as low and high spatial frequency content, fractal structure, device patterns, and defects is feasible in the system. Creative software techniques and analysis can be engaged to deduce these structures in a single-pass scanning mode.

1.2.2 Optical Sensing for Scatterometry

A variety of optical sensors are commercially available to industry, operating at high frequency with wide dynamic response. These can be utilized to track position, and measure displacement and motion, producing a signal, which is proportional to the relative separation between two objects, or the range to a moving target surface; the sensor geometry and mounting configuration identify the related coordinate system, and the illumination to the sensor can be a directed incident beam or by triangulation off a target surface. Sensors are available as single element units or as one- or two-dimensional multi-element arrays, and require mechanical coupling to the target. The sensing element could be a PIN photodiode, or a CCD or CMOS sensor array, or a high gain PMT or APD, and may require external bias and amplification, with the additional possibility of electronic noise. Optical sensing, having advantages of a smaller footprint, improved performance, higher resolution, lower noise, higher speed, simple calibration, stable low-drift sensitivity, long-term reliability, lower cost, and ease of integration, circumvents the limitations of various other non-contact technologies (e.g. charge and current probes) in measurements of planar and contour geometries.

1.2.2.1 An Optical Position Sensor

The optical sensor consists of a planar semiconductor photosensitive active area with contact pairs oriented laterally, as shown in Fig. 1.4, illuminated by a collimated or focused laser beam. The sensor identifies the position of the maximum beam power as follows: radiation is absorbed in the sensor, producing excitons which diffuse radially away from the injection point, electrons diffusing at higher speeds than holes. A steady-state charge distribution centered at the point of injection is established [8]. The cascading structure of the ambipolar electric field is generated from the inequality in the quasi-static diffusion distribution geometries of high-mobility electrons in the conduction band, and low-mobility holes in the valence band. This field has value zero at the point of injection, and increases radially to its maximum value, progressing

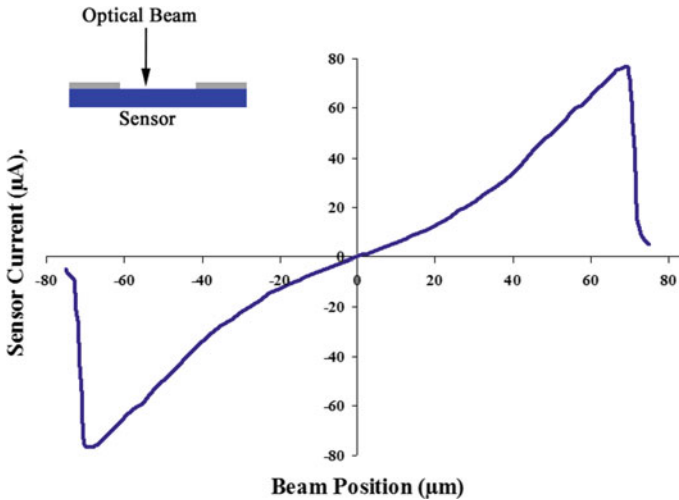


Fig. 1.4 Response of a displacement sensor to the dwell position of an optical beam as it traverses the active area between the contacts [10]

to its secondary null at the occurrence of electron-hole recombination. Prior to this, if this field encounters, interacts, and modulates intrinsic fields within the depletion regions at the contacts, a beam position weighted self-bias is induced at each contact. A unidirectional current flow is stimulated that on average is uniquely proportional to the position of the injection point of the optical beam relative to the contacts.

The sensor in this passive mode is self-deconvolved and is thus virtually noiseless, since the internal currents in the absence of optical stimulus are directed symmetrically opposite at the contacts. An asymmetric beam profile, or a geometrically irregular power profile, stimulates the sensor to identify the position of average-maximum power within the beam, which could be utilized in applications concerning textured or patterned geometries. The sensor signal is always proportional to the position of this “weighted maximum” of the optical beam within the contacts, to the input optical power, and finally to the relative position between the plane of the active surface and the source, which produces resolvable 3D-sensitivity for a multiplicity of solutions in optical systems, such as focusing, wavefront analysis, and beam steering and stabilization. The output signal can be increased to near saturation to circumvent amplifier input noise; the ultimate limit is shot noise in the optical source, which to some effect can be minimized by implementing thermal stabilization of the source, lock-in processing of the sensor signal, and feedback control [9].

1.2.2.2 Static Displacement and Dynamic Motion Sensitivity

An optical sensor was utilized to track the position of a laser beam, and measure the motion of this beam within the contacts. A HeNe laser source (1 mW, 633 nm) was

collimated into the entrance aperture of a 10X microscope objective and focused to a spot of $\sim 5\ \mu\text{m}$ diameter. The active photosensitive area of the sensor was $150\text{-}\mu\text{m}$, and it was attached to a commercially available computer-controlled micropositioning stage. The stage was driven linearly in 100-nm steps as the voltage across it was gradually increased, sweeping the active area of the sensor across the stationary optical beam from contact to contact. The short-circuit current from the sensor was measured directly on a picoammeter, and the digitised value was registered on a data acquisition system.

The positional variation of the signal produced by the sensor in relation to the beam location as it traverses between the contacts is shown in Fig. 1.4. The sensor signal has a precise null at the geometric centre of the active area, and maximises in bipolar fashion at the contact edges. The signal is piecewise linear in symmetric sensor zones, beginning as the beam transits from the shadow of one contact across the contact edge to the active area (first zone), next from proximity to the contact edge toward the central area (second zone), and finally within the central area of the sensor (third zone). This piecewise behaviour is associated with several operational factors including non-uniform sensor illumination by a gaussian beam profile, the shielding of the optical beam as it traverses the contact, and non-linear interaction between the ambipolar field distribution and the internal fields at the contacts. The sensitivity particularly at the contact peripheries is linear and extremely high, providing high resolution of beam displacements within a limited scanning range of a few μm .

Displacement references are readily available for calibration in this configuration. The manufacturer of the stage provides an interferometer-tested specification for its displacement per applied voltage. The sensor itself has a fixed contact spacing, and a corresponding dimension for its lateral active area, which could also serve as a rigid second reference.

The vibratory photoresponse of the sensor to an oscillating beam within the contacts was evaluated utilizing a lock-in technique to explore minimum amplitude sensitivity, and to estimate the dynamic noise floor. In this experiment, the laser beam was coupled into optical fibre, and the fibre was fixed within a narrow tube piezoceramic calibrated at 3 nm/V displacement to the lattice of boronated graphite. A sensor with a $50\text{-}\mu\text{m}$ active area was positioned within the optical beam in the near field, utilizing a mechanical translational stage. The spot size was estimated less than $30\text{-}\mu\text{m}$ full width half maximum. The lock-in amplifier was set to output a 4.8-kHz frequency at 80-mV amplitude signal to the piezo. This excites the piezo and the coupled beam spot to a vibration of approximately 0.25 nm over the photosensitive area of the sensor. The optical excitation and the electrical sensor response are shown in Fig. 1.5. The signal from the sensor was monitored and measured by the lock-in at gain 10^5 . The electronic noise is approximately 20% of the sensor signal, implying a lower limit of sensitivity of approximately 0.05 nm, significantly below the lattice constant of Si. This data suggests that 1-nm beam oscillations could easily be measured in this system.

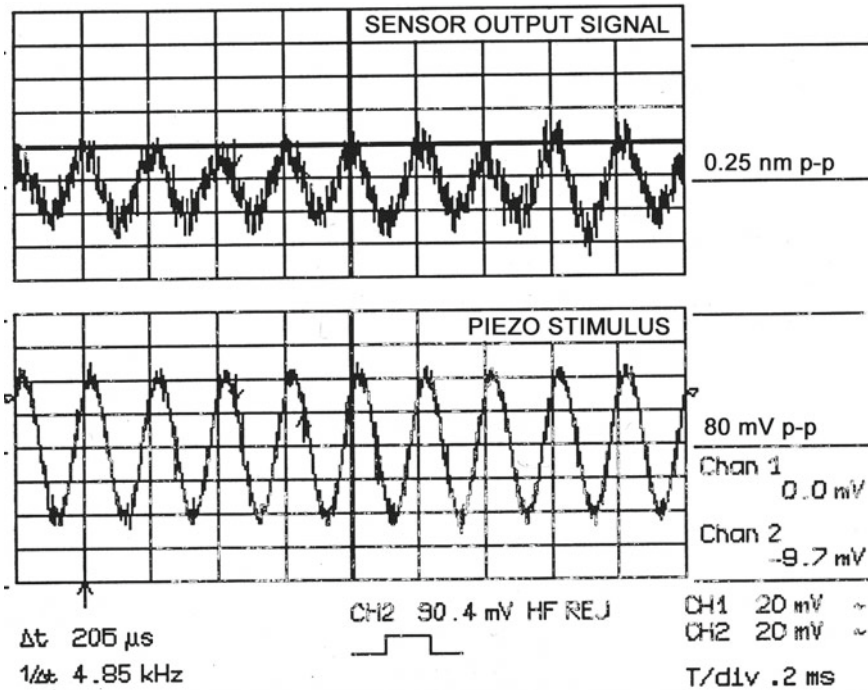


Fig. 1.5 Dynamic sensor response to a 4.85-kHz optical beam oscillation [10]

1.2.3 Laser Triangulation Sensing

The optical scatter and sensing technology described in the previous sections can be engaged to monitor and to evaluate industrial assets, operating logistics, optical and semiconductor fabrication processes, and finished products. The measurements extend from tool parameters to product formation processes, which include tracking cutter position and contour, sensing the motion of rotary, arcuate, and linear components, thermal deformation, spindle imbalance, and comprehensive examination of the machining operation. The complex inspection of diamond cutting tool edges and surfaces required to fabricate optical mirrors, and the sub-wavelength imaging of precision-machined surfaces, defects, and thin films are also feasible.

The metrology and imaging functions can be integrated into a single multi-channel sensor-probe, consisting primarily of a stable laser source, sensors, a triaxial scanning piezo stage with range of 50 μ m \times 50 μ m \times 50 μ m (100-nm pixel resolution), ultra-low noise electronics, data acquisition, and integrated with operating software to enable motion control, signal processing, data analysis, and graphics. A 785-nm laser diode is utilized as the optical source. The sensors and the source are oriented and mounted adjacent, in a planar configuration on a base, with adjustments to facilitate the applications and the measurements detailed below. The beam spot

in the intermediate field zone is under 0.1-mm diameter, and can be scaled to the application, utilizing the normal axis of the stage.

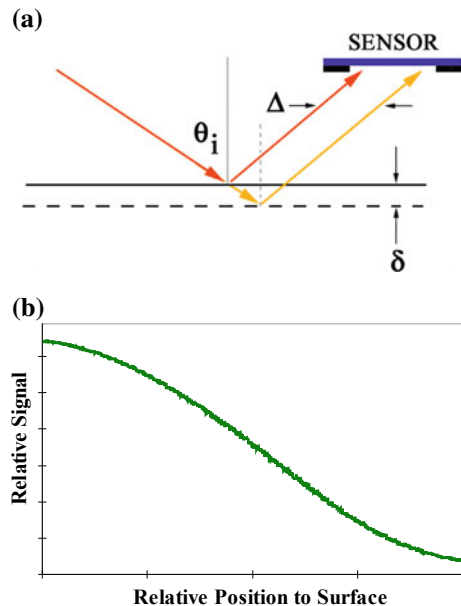
The in-plane alignment of a sensor to the optical input beam is as shown in Fig. 1.6a. This geometry is referred to as laser triangulation, wherein the optical source and various sensors are aligned at angles such that the source beam reflects from the target surface to a specular sensor, forming a plane of incidence orthogonal to the test surface. Additional sensors may also be installed to measure scatter as required. This versatile non-contact configuration is optimised for various sensing and imaging applications, and may be utilized for a wide variety of target materials.

The displacement and motion sensitivity of this sensor probe can be described with simple geometrical arguments. Let the plane of the target surface move a parallel distance δ away from its reference position (Fig. 1.6a). The optical beam is offset to a new position on the test surface, translating the reflected beam laterally across the active area of the sensor by a dimension, Δ . The sensor responds by producing a signal proportional to this displacement, according to Fig. 1.4. If the triangulation angle θ_i is set at $\pi/4$, the displacement of the optical beam is levered by a factor of 2δ , with approximate equivalence occurring at an angle $\theta_i \sim \pi/7$. This is a practical lower limit to offset the plane of the sensor beyond the Wiener wave distribution at the surface, and thereby circumvent potential interference in near-field measurements.

The vertical response (in normal orientation) of this triangular beam configuration over a perpendicular scan of $50\ \mu\text{m}$ toward a test surface is shown in Fig. 1.6b. The response of the sensor probe within the central area of the scan is linear, rendering it very useful for various sensing and ranging applications, which we cover in detail below.

Fig. 1.6 a Optical lever: contact-free sensing with noiseless gain [10].

b Displacement response Δ of the triangulated sensor-beam arrangement to ramp δ perpendicular to the surface plane, 50- μm total [10]



This high-sensitivity optical probe, synchronized with lateral two-dimensional piezo scanning capability, comprises a non-invasive imaging system. The scatter signal is measured by positioning one or more separate sensors apart from the specular beam, and adjusting the altitudinal and azimuthal orientations to detect spatial roughness or defect types, as shown in Figs. 1.1 and 1.3, and described in Sect. 1.2.1. The system utilizes unique software algorithms to operate scanning functions, adjust resolutions, process various signals, reduce noise, and enhance the sensor response to accurately measure scatter contrast. The versatile usage of this integrated technology is explored with evaluations of various machine tool operating parameters and examinations of finished surfaces, geometric artefacts (edges), defects, and thin film layers in 3D measurements, which are presented in the following sections.

1.3 Fabrication of Optical and Semiconductor Surfaces

Optical scatter metrology combined with adjustable triangulation forms an innovative tool for comprehensive non-invasive inspection in fabrication processes for precision optics and semiconductor wafers [10]. The key features of the technology described in the preceding section are its capabilities of contact-free non-destructive evaluation, high resolution, ease of use, and simple integration. The software is designed to enhance sensing flexibility, improve defect-to-background contrast, identify known targets, and provide instantaneous viewable results.

We now utilize these advantages and versatility to perform critical measurements on a standard machine tool, and investigate its operation together with a comprehensive analysis of finished products. The limitations of this fabrication system are explored by a careful review of the data. The potential for true nanometre-accurate fabrication of optics and semiconductors is explored, together with the possibility for improvements in product quality and yield [10].

1.3.1 *Spindle: Torque Offset and Thermal Stability*

Precision machining requires careful analysis and consideration of several factors involving tool hardware, the cutter, lubrication, the environment, and overall system stability. The tools are designed as integrated systems and require gravitational levelling combined with vibration isolation, rotational and translational stability, strict temperature control, and a clean room environment to meet optical surface tolerances. In order to achieve this high level of accuracy, the tool must operate with pressurized fluid-lubricated (hydrostatic) bearings supporting the rotating spindle and cutter tracking to maintain uniform low-error fabrication processes. The tools are provided by the manufacturer with specific operational procedures and techniques to achieve their narrow specifications. It is customary and beneficial for the end user or

operator to frequently measure and monitor system performance to accomplish repeatable product integrity and yield.

Axial, radial, and translational error motion are measured schematically as shown in Fig. 1.7. A cylindrical work piece is clamped to the hydraulically supported spindle utilizing a vacuum or mechanical chuck to ensure rigid coupling. In a preliminary procedure, positioning error is reduced by implementing a static preadjustment of the relative clamped position of the blank within the chuck to align its surfaces to the spindle and to the cutter. In a subsequent dynamic procedure, the cylindrical surface and the front surface of the blank are then pre-machined (few μm of progressive removal) to true workpiece geometry in circularity to the rotary motion of the spindle, and surface planarity to the linear motion of the cutting tool.

The cutting tool is mounted on a holder stage with translational and rotational adjustments over linear and angular degrees of freedom, and to which the sensor probe is also coupled with nanopositioning capability. Thermal and rotational characteristics of the spindle-bearing assembly, and translational accuracy of the cutting tool were examined in the triangulation-sensing configuration (Fig. 1.6a, b) to study the overall performance of the machine tool and the sensor. The sensor was first positioned on axis, centred on the spindle, which was set to operate in free running mode (100 rpm for approximately 30 min) to thermally stabilize the system. The axial error motion on this unit in this operational phase was approximately $\pm 3 \mu\text{m}$ (Fig. 1.8). This error in the rotating spindle originates in paraxial torque-vector variation, which dynamically repositions the spindle, causing weakly damped axial, and (to a lesser extent) radial instability within the bearing encasement.

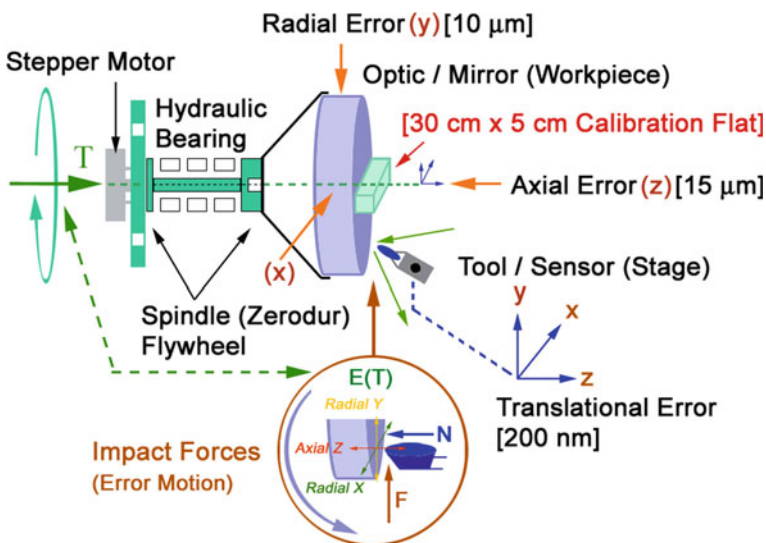


Fig. 1.7 Tool schematic showing spindle, error measurements on the workpiece, and translational error (measured on the calibration flat) of the cutting tool holder stage [10]

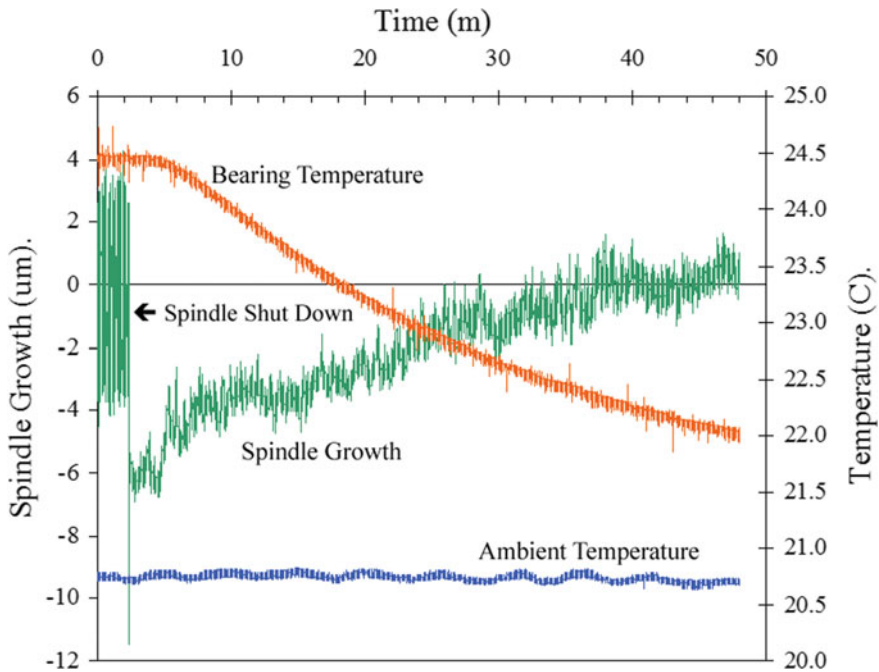


Fig. 1.8 Analysis of thermal and rotational instability over long duration in a precision spindle [10]

Once thermal equilibrium is established in the initial rotational phase, the spindle is immediately brought to rest. The step transition in the data, occurring at time ~3 min reflects the rapid deceleration of the spindle in the transition to its stationary state. Static torque-free axial readjustment of the spindle takes place within the bearing. The dynamic-to-static positional offset is 6 μm , approximately the same as the rotary axial error. The high-frequency noise of less than 1 μm in the data is a measure of stray relative vibrations between the resting spindle and the fixed sensor mounted on the tool stage. The spindle now rests passively in uniform fluid suspension, balanced in axial and radial orientations within the bearing. The bearing encasement temperature was 24.5 $^{\circ}\text{C}$ during the steady-state rotational phase. It stabilizes and compresses incrementally to a temperature of 22 $^{\circ}\text{C}$, over the period of an hour after power down. During this time, the spindle extends back to its average position in the rotational phase, with the gradual transfer of thermal energy to it from the bearing assembly. This thermal growth error at approximately 6 μm , matches both the dynamic torque-axial error and the static offset, while the ambient air temperature remains constant at 20.7 $^{\circ}\text{C}$ during this evaluation.

1.3.2 Tool-Holder Stage: Positional Accuracy and Backlash

The cutting tool is firmly clamped to a multi-axis stage with Cartesian and angular degrees of freedom to enable the manufacture of free-form geometries extending from smooth planar surfaces to contoured aspheric mirrors. Triaxial static Cartesian accuracy of the stage was evaluated by clamping a 30 cm \times 5 cm polished calibration flat to the spindle chuck (Fig. 1.7), mounting a sensor to the tool holder, and recording the signal as the tool holder is programmed to perform multiple translational functions.

In the first measurement, the positional accuracy and hysteresis of the tool stage were evaluated parallel to the spindle axis in 125-nm steps as follows: a 20-step sequence, retracted away from the flat surface, followed by 40 steps in the direction of the flat surface, and finally a 20-step retreat from the flat, moving the probe back to its start position. The results are shown in Fig. 1.9, with an expanded view (inset) of each command step over a span of 2 μm . The accuracy of the stage translation is well within normal tolerances for a mechanical system with error margins of a few 100's nm. There are roughly 3+ units of translational backlash, or less than 500 nm error, as positional-step instructions are sent to the tool. The roundtrip tool-point travel from its original resting position through the 80-step programmed sequence is 10- μm , while the round-trip hysteresetic error is less than 30 nm.

We next performed static translational measurements perpendicular to the spindle axis (i.e. oriented along the radial axis). The sensor was first scanned across the 30-cm lateral dimension of the calibration flat, offset 2 μm beneath the original track, and then scanned back to its initial position. The bi-directional scan data displayed in Fig. 1.10 indicates error margin in the 100-nm range, and equivalent repeatable accuracy of the tool positioning system. The shape error of the flat is also observable

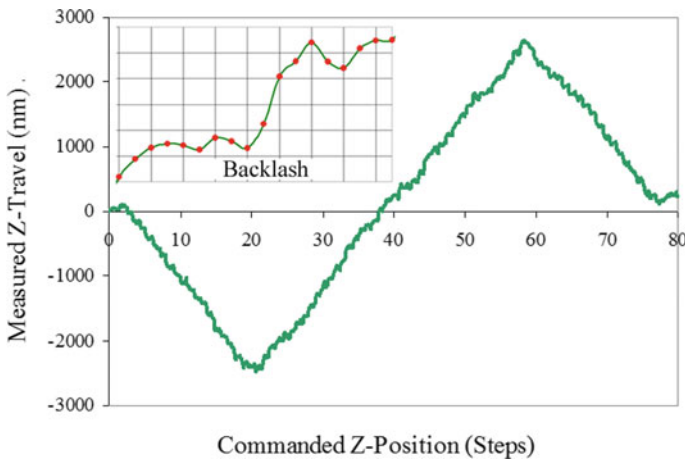


Fig. 1.9 Sensor data for an 80 \times 125-nm forward-and-retracted step pattern, parallel to the spindle. The expanded view shows \sim 3 points of backlash, and the cumulative error is $<$ 500 nm [10]

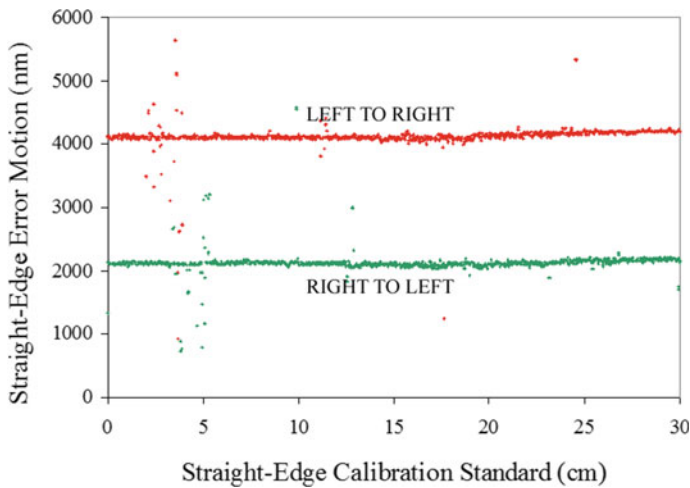


Fig. 1.10 Radial scan data, in bi-directional translation across the calibration flat. The error is ~ 200 nm. Note the divergence of the data at ~ 18 cm, from shape error [10]

as an upward trend to a 200 nm offset in the data commencing at 18 cm. Contamination of the reflective surface of the flat is evident in the spurious error signals (Rayleigh scatter) particularly in the preliminary 5 cm of the scan. Cumulative assessment of these measurements indicates that the triaxial x-y-z tool-tip translation is capable of performing machining accuracy, with minimal hysteresis, to less than 500 nm.

1.3.3 Spindle: Dynamic Axial and Radial Instability

We migrate now to an experimental analysis of the dynamic motion of the tool where the spindle is operated free running for a long duration, and allowed to attain thermal stability. Each axial and radial measurement was acquired on the spindle over a 5-min period of steady state rotation at speeds of 100, 500, and 1000 rpm for error-motion analysis.

Figure 1.11 contains corresponding data on the axial rotation of the spindle. There appears to be significant high-frequency vibration of the spindle, with periodic and aperiodic content in the 100-rpm data. The amplitude of this noise is approximately $6 \mu\text{m}$. This error is superimposed on sub-Hz frequency axial instability of amplitude $15 \mu\text{m}$. At rotational speed of 500 rpm, the spindle is inertially balanced and the rapid error noise is damped to below $1 \mu\text{m}$, however the $15\text{-}\mu\text{m}$ gradual instability persists. The same is also observed when the spindle rotates at 1000 rpm, but the high frequency error motion now registers at under 500 nm, indicating instantaneous axial balance.

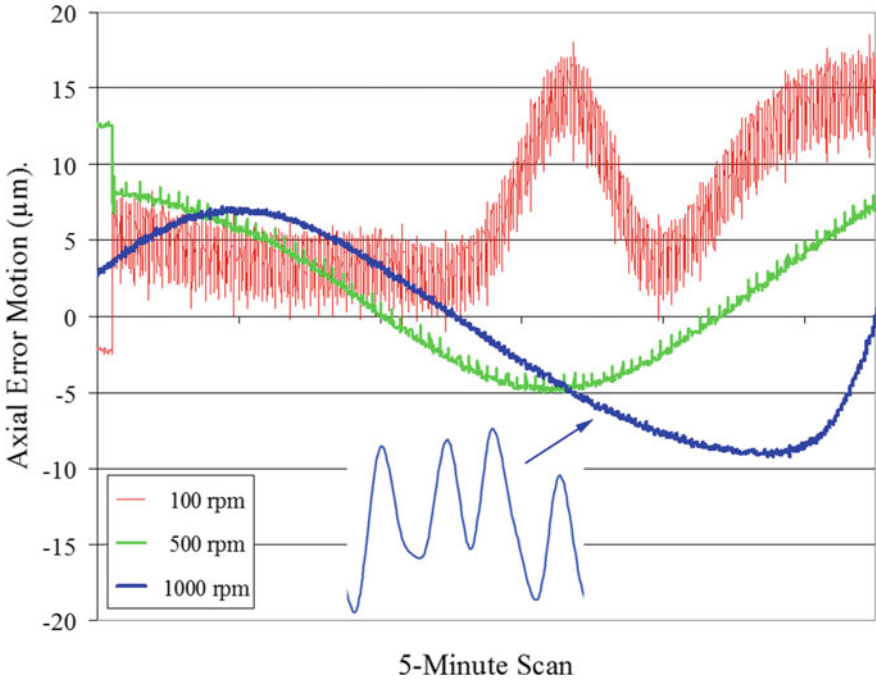


Fig. 1.11 Measurements of axial error motion at 100, 500, and 1000 rpm made over a 5-min duration. Expanded view of 1000 rpm axial data indicates under 500 nm noise [10]

Spindle noise error is observed as expected in the radial orientation as well, and is lower than the corresponding axial noise at all speeds (Fig. 1.12). This is anticipated and intuitive since the spindle is by design supported radially within the bearing. The high-frequency periodic and aperiodic radial fluctuations at 100 rpm are at a reduced amplitude of 3-µm, in combination with a sub-Hz imbalance error of less than 10 µm (including high-frequency elements), implying existence of a much lower level of radial torque error in the bearing. Inertial stability of the spindle at 500 rpm improves to below 1 µm. The sub-Hz radial instability in this case is under 5 µm, corresponding to ~30% of the equivalent axial error amplitude. At 1000 rpm, the spindle has approximately 250 nm of high frequency noise, and 5 µm of radial imbalance.

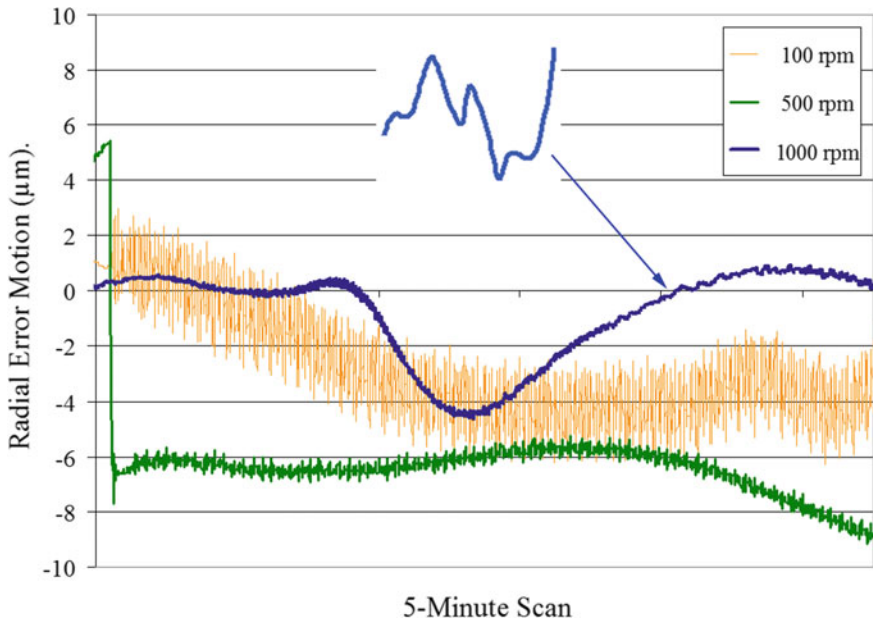


Fig. 1.12 Measurements of radial error motion at 100, 500, and 1000 rpm made over a 5-min duration. Expanded view of 1000-rpm radial data shows <250-nm error [10]

1.3.4 Spindle: Error Motion Analysis for Product Formation

The measurements presented in Sects. 1.3.1–1.3.3 provide a fundamental analysis of each moving mechanical component of a typical production system for optics and semiconductors wafers. Each plot contains noise components, which could in principle be decoded and traced to a corresponding source, following which remedial adjustments may be designed into the system either in the hardware or by software. Similar analysis of data from various tool components, with parametric adjustments (e.g. speed, feed rate, removal depth, etc.) can then be utilized to understand the impact of system error on the finished surface of an optic or a planar wafer. This can serve as a guide to provide specific metrics of tool production capabilities, and to select optimized points of operation for repeatable nanometre accurate quality and high manufacturing yield.

We discuss with some emphasis the role of spindle instability in the introduction of surface roughness and shape error in product formation. Figures 1.11 and 1.12 contain dynamic measurements on axial and radial spindle stability over a wide range of speeds. Error motion of 1-µm amplitude was observed at short duration but highly problematic is the sub-Hz 15-µm instability over long durations utilized for fabrication of large surface areas; the periodic time constants are more consistent with mechanical factors as opposed to very long duration asymptotic thermal processes as shown in Fig. 1.8. The sources for this noise could be independent or

synchronized combinations of factors such as drive-motor speed fluctuations which induce torque variation, spindle alignment and inertial imbalance, spindle material and shape inaccuracies, suspension instability caused by non-uniform bearing fluid pressure, encasement and fluid acoustics and resonances, stray frictional forces and sporadic energy transfer to the encasement, insufficient broadband vibration isolation, gravitational tool imbalance, insufficient heat dissipation with internal thermal gradients, and various unknown stray factors. The removal of material under the influence of all these factors could result in progressively worsening inertial imbalance and further deviation from nanometer surface and shape quality. The instability and precession at a given point within the spindle is mechanically levered (amplified) to larger amplitude at the work surface, pivoting at a point on the inertial spindle axis, within the bearing. Instinctively, motion error should ideally be sensed on the workpiece to fully comprehend and analyse the errors within the spindle and the entire mechanical system at large, and to enforce some means of nullification.

These enumerated factors could serve as research areas in a feasibility study for the future goal of true nanometre production. Machining systems for optical fabrication are typically designed with flywheels, which have passive adjustment to rebalance the spindle and to evanesce non-circular motion. Zerodur, which purportedly is a zero thermal expansion material over a short temperature range, is utilized to manufacture spindles in precision systems. Yet another possibility is to devise a computerized sensor-actuated adjustment and damping system to dynamically smooth out large μm -level instability to below 100 nm.

Single point diamond turning (SPDT) is considered state-of-the-art by which an optical surface is prepared for high-resolution imaging. The tool contacts the workpiece surface in a machining area defined by the edge radius of the tool ($\sim 1\text{--}2\ \mu\text{m}$) and the cutting arc ($< 10\ \mu\text{m}$); we refer to this as the contact region of the cutting tool. The removal depth is also very shallow ($\sim 1\ \mu\text{m}$) to maintain the removal process in the damage-free ductile mode. The expanded view of this process is shown in Fig. 1.7. The tool exerts normal force (N) with associated friction (F) during the machining process. Figures 1.11 and 1.12 indicate that the maximum axial and radial instabilities are on the scale of $15\ \mu\text{m}$ and $10\ \mu\text{m}$, respectively. The cumulative instability in the spindle produces a counter-normal impact error force (E), sourced by torque error (T). The forces E, F, & N combine in the machining area to generate high friction, anomalous heating (diamond is an excellent conductor of heat), and acoustic shock in the SPDT process. The vibrational error of the workpiece surface is thus transmitted directly from the machining area to the contact region of the cutting tool, where it produces proportional periodic and aperiodic impact forces combined with thermal deformation of the tool edge. This significant energy transfer from the machining area to the cutting tool point-of-contact, gradually erodes the edge profile and its contour, and limits its life span. The degradation of the edge profile of a cutting tool is examined in Sect. 1.3.5. Note that the translational error of the tool-holder stage, and the rapid spindle error motions in the range 100–500 nm are significantly smaller, and have reduced impact on the tool edge.

We first examine the impact of these errors on the process of machining a flat mirror surface. If we utilize a fresh cutting tool, the axial instability (Fig. 1.11)

of the spindle would introduce roughness of equivalent amplitude (depth) in the machined surface. The radial error (Fig. 1.12) would modulate the spread or spatial frequency of this roughness in the plane of the surface. The positioning error in the tool stage is on the order of 100 nm (Figs. 1.9 and 1.10), and produces finer roughness texture. Clearly, it is highly desirable that all sources of error motion and positional inaccuracy are minimized to achieve true nanometre surface quality and produce true planar flatness.

The machining of a curved mirror surface or a lens to a specified contour is a far more complex challenge. The dynamic axial and radial instability (Figs. 1.11 and 1.12) would imprint proportional roughness and shape error in the optical manifold. As in the case of a planar surface, the axial vibration machines proportional roughness (depth) in a radial distribution on the finished surface. The radial instability would spatially modulate this roughness, and in addition introduce local error in the shape. The error in the tool holder is also a contributing factor since it is programmed to position the cutter to implement the arcuate surface topology. This error, while minor, is inherent in both the Cartesian and the angular positioning of the cutting tool required to produce the desired shape. So we see the possibility of significant surface and shape anomalies when all error sources act in tandem.

The data in Figs. 1.11 and 1.12 indicate that a motorized rotating system with a hydraulic spindle has an operating threshold below which there is significant inertial instability and multiple resonance frequencies in the tool, and partial aperiodic content of unknown origin. The cutting operation is improved if the rotating speed is increased well beyond this threshold, where all high frequency instabilities evanesce. This could pose a limitation for materials like Ti, which require low machining speeds. In general, to quantify product errors, it is beneficial to mount the cutting tool with a sensor in identical arcuate positions (Fig. 1.7). It would then be feasible to machine the surface, and perform a follow-up measurement of its contour to examine its shape accuracy, within tolerances established by the error noise of the mechanical system. The fabrication of precision planar and curved surfaces in optics and semiconductors requires detailed analysis of a variety of parameters, and experimental measurement of tool performance at various speeds to optimize operational parameters for nanometre quality products.

Taking a step back, the measurements and the analysis potentially form a basis in the precision industry for the design and manufacture of nanometre accurate systems. It offers guidelines for refining the design of balanced spindles and bearings suspension (e.g. with pattern geometry for intra-surface laminar fluid flow), improving the translational accuracy of tool-holder stages, implementing stability in dynamic processes, absorbing unwanted resonances, effecting vibration isolation, and minimizing unwanted thermal gradients. The parameters of depth of cut and feed rate, coolant dispensation, spindle speed, bearing fluid pressure, component temperature, and machine tool stability, can be optimized in the manufacturing process for nanometre surface roughness and shape accuracy with improved abundant yield.

1.3.5 Cutting Tool: Impact of Error Motion

The discussion in Sects. 1.3.3 and 1.3.4 focused on the mechanical instability of the work piece traced back to spindle errors, and lesser positional errors of the tool holder stage. The work surface can be seen as essentially pivoting on the cutting point of the tool in SPDT, with a steady transfer of kinetic impact and energy between the two components. The cutting tool absorbs this energy at the microscopic contact point, which gradually erodes, modifies, and deteriorates its profile. SPDT cutter specifications are as follows: transverse edge radius at the cutting point, contour radius and roundness, edge roughness, cutter angle, rake face planarity and rake angle.

Contour condition is an excellent indication of usage and utility. An image of the edge contour is particularly useful to identify degradation of an SPDT cutter, which correlates directly to machined surface quality. This image would in addition contain information on the rake face as well as the adjacent flank. The erosion of the edge is directly related to duration of use, depth of removal, spindle instability and stage inaccuracy, friction and associated thermal gradients, cooling technique, rotational speed and feed rate, material and surface properties, and the crystalline perfection of the diamond tool.

A comparative analysis of two SPDT cutter edges was carried out to examine the differences between a factory fresh unit, and another unit, post-utilization in the formation of multiple polycarbonate lenses. The tools had identical specifications i.e. zero rake angles, $\sim 2\text{-}\mu\text{m}$ edge radius, cutter angle of 60° , and 0.2-mm contour radius. These commercially available cutters are fabricated by control polishing of the single-crystal crystal diamond blank with progressively finer natural diamond particles. The blanks typically have crystalline defect inclusions, and sub-surface damage in the fabrication process is unavoidable.

The cutting point of the new SPDT tool had excellent edge quality and a clean rake face and flank, as shown in the minor image of Fig. 1.13. The contour radius is measured exactly at 0.2 mm in agreement with the manufacturer specification. There were no visible edge defects and no observable edge roughness. The rake also appears to be planar, devoid of any manufacturing defects such as polishing texture originating at the edge.

The second tool was installed in a manufacturing unit to produce dozens of eye-wear lenses. This tool has multiple artefacts of interest as seen in Fig. 1.13. We observed an uneven edge-contour profile in the rake face encompassing the flank, with deep furrows spanning down the flank. These modulations range from 1 to $10\ \mu\text{m}$ in spatial spread and amplitude, and correlate closely with equivalent multi- μm axial and radial sub-Hz instability in the spindle.

In SPDT, the combined system noise error produces equivalent high impact forces (Fig. 1.7: $E + F + N$), transferred from the machining area of the work surface to the edge of the cutting tool, within its microscopic region of contact (Fig. 1.7, inset). The spindle instability described in Sects. 1.3.3 and 1.3.4 produces friction, acoustic shock, and local thermal gradients during fabrication to discompose the cutting tool

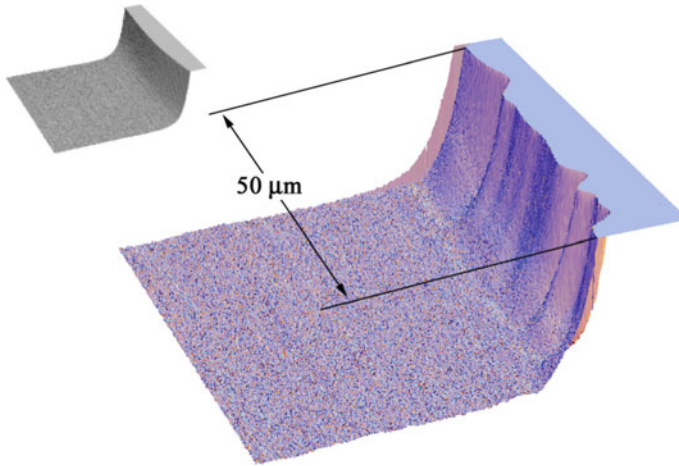


Fig. 1.13 Comparison of the cutting contour of diamond tools. The fresh tool (left) has an accurate noise free edge. The contour of a used diamond tool shows wear grooves in the range of 1–10 μm , perhaps correlating with spindle error [10]

edge at defect sites, nanoroughness structure, and subsurface damage. We note here that thermal effects within the spindle-bearing assembly shown in Fig. 1.8 also do play a role, albeit minor in comparison to tool error motion. A gradual removal of diamond material from the edge occurs concurrently while the work surface is being shaped. From our measurements we observe that the contour-flank wear patterns become progressively deeper with usage, effectively increasing the transverse radius of the cutting edge thus reducing its sharpness for removal.

Intuitively, a cutter with a deformed contour and a larger edge radius will perform very poorly in a precision environment. The modified cutter wear structure will be transferred to the machined surface, fabricating a product with unacceptable quality, having visible optical defects and deformed shape. Extending the discussion of Sects. 1.3.3 and 1.3.4 to the measurements and the observations in this section, it is conclusive that a machined surface fabricated utilizing this particular diamond tool, with the system operating at low rpm, could not be considered optical.

1.3.6 *Product Analysis: Surface, Defect, Thin Film Inspection*

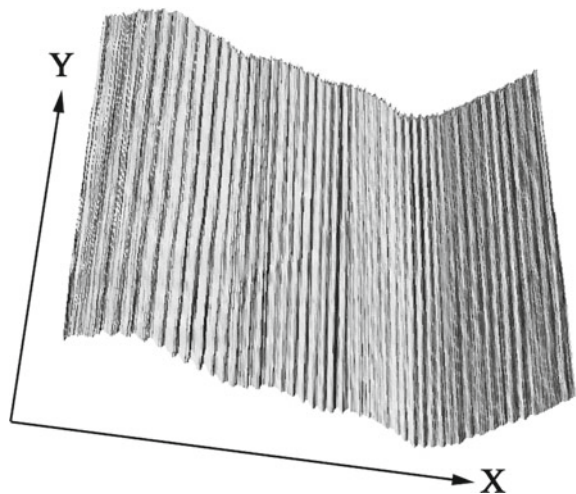
The comprehensive evaluation of precision surfaces and the detection of nanoscale defects on optics and semiconductor wafers, single-layer or multi-layer thin film coatings, and photoresist patterned substrates is frequently carried out by non-invasive rapid scanning optical scatterometry. A scatterometer system to perform this evaluation is shown in Fig. 1.1. It consists of detectors oriented to the test surface, and

configured with optics and electronics to measure radiative scatter in a solid angle, in response to an input optical beam at a fixed angle of incidence and a defined polarization state.

As we discussed earlier, smooth surfaces produce uniform low intensity scatter haze, requiring that the detector signal be amplified by low noise, high gain electronics, and coupled to spatial filters. The input optical entry system could include a compensator to adjust the polarization to a particular orientation and a coherence diffuser, thus improving feature-to-background sensitivity, despite operating at a reduced power. Our system includes a 3D nanometre scanning stage combined with software processing of scatter data to produce contrast-enhanced images of surface roughness, defects, and sub-surface damage.

The scatter image of a diamond turned 1-cm² Si-wafer surface in a 50 $\mu\text{m} \times 50 \mu\text{m}$ field is shown in Fig. 1.14. The fabrication parameters are as follows: the spindle rotation was set at 1000 rpm, and a 1.5-mm diamond cutter (unknown edge profile) was set at a feed rate of 3- $\mu\text{m}/\text{rev}$, with 1.3- μm material-removal depth. This process machined indentations on the Si surface, conforming to the feed, in a pattern of lines. The surface is undeformed in orientation along the y-axis (radial axis), however along the x-axis (axial axis) the surface is distinctly non-planar and contoured with structure of about 10–25 μm dimensions. We observe that this dimensional range is in close agreement with the data on axial spindle error, shown in Fig. 1.11. The axial error being more pronounced makes the surface non-planar to a larger degree. The smaller radial error (Fig. 1.12) does also leave a shallower imprint, which is observed in the image in the central area near the right border (see discussion in Sect. 1.3.4), and it is oriented diagonally across the surface. We note here with some emphasis that the fabrication of a large optical surface on this machine, with this particular cutter, would imprint μm -size irregularities in the surface and the shape.

Fig. 1.14 Diamond-turned Si surface (50 $\mu\text{m} \times 50 \mu\text{m}$ scan). The cutting tool indentations appear as grating lines parallel to the y-axis. The non-planar shape error is a result of axial and radial spindle instability [10]



Following a rough surface measurement, the scatterometer system was utilized to examine planar surfaces polished to lattice smoothness. The data yielded a simple flat field, and microstructure and defects were not observed. The specular beam had a virtually replicated beam profile as the incident beam, with minimal Lambertian scatter haze. This surface data was verified by performing a high-resolution atomic force microscope (AFM) topographical scan in the identical vicinity of the scatter measurement.

Large optics and Si wafers are frequently manufactured with a random distribution of surface defects. The scatter features of such defects are of great interest to the respective industries so that they can easily be identified and their location noted. We examined a highly polished quartz flat, scanning the surface with the same parametric settings as previously; the data is shown in Fig. 1.15. There are several artefacts, including low spatial frequency roughness of 10- μm period, oriented perpendicular to the x-axis near the origin. In the rear area of the scan, the surface structure is 5–10 μm in spatial spread, along both axes. The periphery of a second defect in observed in the far corner. The central defect is a crystal originated pit or COP, quite commonly identified on polished Si wafers. It is approximately 15 μm in diameter, and the scatter image shows that the rim-perimeter does not conform to the geometry of a circle in the surface plane; it has shallow wavelike hemi-toroidal structure. An outer concentric ring appears slightly more circular with minor vertical peripheral projections in the range of a few micrometers, which perhaps could be a secondary diffraction maximum of the COP rim. Note that this level of edge resolution is not observed in the image of a similar defect on the same quartz flat produced by AFM, to the right. The AFM image does show some coarse surface granularity, and a distribution of minor pits in the surface. The scatter image of the defect also appears to have a minor internal bisecting bridge, spanning the pit, oriented along the y-axis. While scatter can produce an image of the internal structure, AFM displays the

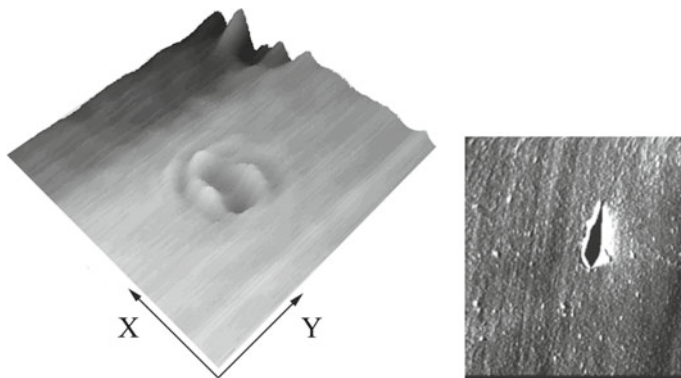


Fig. 1.15 COP defect on a polished quartz surface. The surface structure is composed of wave like patterns with particular orientation. An AFM image of a similar defect and surface structure on the same sample is shown on the right. Note that the scatterometer can image the interior of the defect, while this appears as a dark area in the AFM scan [10]

interior of the pit as a simple lateral dark space with no transverse (vertical) edge contour definition.

The final scatter measurement is of a $75\ \mu\text{m} \times 75\ \mu\text{m}$ bi-layer thin film composed of 50-nm Cr and 150-nm Au layers, deposited by e-beam evaporation on a Si wafer. We were interested in measuring the scatter produced by the surface texture of the film, identifying the two metal layers at the edges, and examining the corner. Figure 1.16 is the scanned image, which reveals the layering structure of the film, in greater detail than is possible by standard high-magnification, diffraction-limited, optical microscopy.

The film thickness of 200 nm is at approximately 25% of the wavelength of optical probe beam, and is clearly visible in the image. The Cr-base region of the film has low-reflectivity, in comparison with the highly reflective upper-Au layer, which produces brighter scatter. The top surface of the film consists of randomly oriented fine grain texture in the spatial frequency range of a few micrometers laterally, and a few 10's of nanometres in vertical amplitude, which is less than 3% in dimensional comparison with the sensing wavelength. The bi-layer film deposition process appears to create islands of molecular nucleation, resulting in fractal crystallites of Au on the surface. The surface of the film may also mimic the underlying structure of the Cr film. This data reveals that the scanning scatterometer-triangulation technique has excellent depth sensitivity in its imaging function.

The peripheral film geometry is also revealed in this measurement. A photoresist-mask pattern was utilized in the lithographic lift-off fabrication process. A uniform ridge remains along the upper edge rims of the film, identifying the break between the

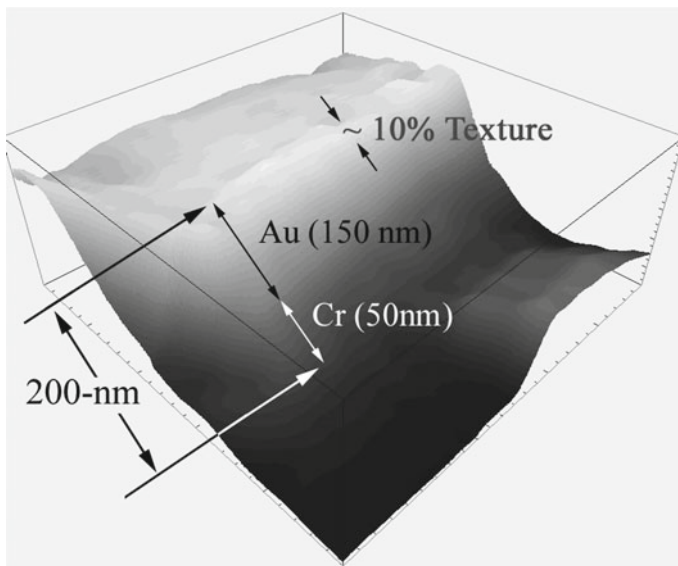


Fig. 1.16 200-nm Cr–Au film on Si. The Cr and Au film layers are identifiable as shown, and the surface of the film appears to be fractal structure formed by nucleation [10]

contact and the extraneous layers. Optics and semiconductor devices utilize dielectric and metal thin films as coatings to perform functions such as high-reflection, anti-reflection, devices, connections, and contacts, whose structure, uniformity, and defect distributions could easily be examined by this scatter technology at nanometre resolution.

1.3.7 Scatterometry: Future Applications in the Environment

Light scatter is ubiquitous in nature, and Lord Rayleigh presented the first analytical formulation of scatter in his brilliant analysis and discussion of the blue sky over a century ago. It is somewhat paradoxical that the industrial age has ushered in some serious challenges to keeping the sky blue in several of the world's larger cities, and their surrounding areas. The twin consequences of atmospheric warming and climate change offer opportunity for some serious thought on how to address the issues of monitoring and reduction of industrial effluents released to the environment, a situation that is unremitting and international.

One of the prime fomenters of the current global situation is the massive proliferation of internal combustion-engine driven automobiles in many parts of the world with no decline in sight and scarce environmentally friendly alternatives. There are just fewer than 1 billion units in ownership worldwide, with about one quarter of these in constant operation at any given time, continuously emanating pollutants and heat. While auto emissions standards are high in some areas they are non-existent in most others. Operation and maintenance provide associated challenges for the effective management of solid, liquid, and gaseous wastes, and heat waste regularly produced within this system, which is supported by a massive infrastructure worldwide, with minimal effort at recycling and conservation.

Sensors based on scatterometry could effectively monitor industrial and automotive effluents like gases and particulates, which pose serious health hazards, particularly to young children. Scatterometry may be combined with various methods of spectroscopy to hone in on, and identify low-density high-risk contaminants.

Scatterometry at selected wavelengths could be utilized in remote sensing applications in areas such as agriculture and irrigation, water resources, forest fires, and protecting natural resources from by-products of the incremental encroachment of industrialization.

Perhaps, the time is ripe for the Nobel Committee or a similar benefactor to commence issuing a notable annual competitive international prize to stimulate creativity, encourage new technology, and promote sea change advances in Environmental Science, to steer international society away from manifold practices detrimental to the global condition, which at present seem only to increase over time. The science of scatterometry is well known and widely available, and effective solutions are pending.

1.4 Conclusions

We have presented optical scatter technology with an emphasis on measurements, applications and analysis in the manufacture of optics and semiconductor wafer surfaces. A sensor-probe was developed by uniting laser triangulation with scatter measurements, utilizing a passive semiconductor device. The sensor-probe was integrated with a 3D nanometre positioning stage. This PC-based system utilizes standard low noise data acquisition components, and multi-function analytical software to interpret data in real time and produce 3D images. It is also configured to provide feedback information to the manufacturing system for adaptive process control. Our results indicate the versatility of this technology for rapid contact-free, non-destructive evaluation of the operation of fabrication equipment for optical and Si-wafer surfaces, and to examine the finished products to accuracy of less than 100 nm, with resolution approaching 10 nm. We explored thermal deformation, spindle torque vibration, tool positioning error and backlash, and axial and radial error motion, and presented an analysis of the impact of these factors on the product surface and shape, and the need to minimize these instabilities to maximize quality and yield. The spindle exhibited sub-Hz axial instability on the order of 15 μm , and equivalent radial instability of $\sim 5 \mu\text{m}$ at operating speeds of 1000 rpm. The low 100-rpm operation contained considerable $5/2 \mu\text{m}$ high-frequency axial/radial error motion; 1000-rpm operation settled to under 500-nm error. A diamond cutting tool utilized in the manufacture of plastic eyewear was examined and had 10- μm edge erosion patterns as a result of these spindle instabilities. A Si wafer machined by SPDT had progressive linear imprints equal to the machining feed rate, and 10–25 μm shape error as a result of axial spindle error; the imprint of transverse radial error was also observed. The comprehensive measurements and analysis presented herewith could assist in the design and development of machine tools for nanometre quality. Some areas of utility include the fabrication of critical components like the spindle and bearing surfaces to high precision, reduction of vibratory noise and thermal creep in the system, and formulation of fabrication processes with the selection of optimal operating points, tailored to the manufacture of optical and semiconductor products with nanometre quality and high yield. A highly polished quartz optic was imaged and a COP defect, its uneven perimeter, and its internal structure were measured. The edges and the surface of a composite 50/150-nm Cr/Au film were scanned, and the Cr and Au films were identified by their differing reflectivities; 20-nm fractal nucleation structure was seen on the surface of the Au film. We foresee applications of scatterometry in addressing multiple challenges in the areas of environmental protection, climate change, the preservation of natural resources, and optimising energy efficiency.

Acknowledgements Financial and/or technical assistance for this work was provided from several sources at various times, including the US Naval Research Laboratory, California Institute of Technology, US Air Force, NASA—Jet Propulsion Laboratory, the Livermore Laboratory, the National Research Council; and is in dedication to Esmeralda and Benjamin Soares. Dr. Brian Limketkai and Dr. Payam Pakzad at Caltech provided helpful experimental assistance in this project. Additional assistance was provided by Dr. Andrew Huntington and Dr. Neil Jones at Caltech.

References

1. R. Azzam, N. Bashara, *Ellipsometry and Polarized Light* (Elsevier, 1999)
2. O. Heavens, *Optical Properties of Thin Solid Films* (Dover, 1955)
3. E. Church, H. Jenkinson, J. Zavada, Relationship between surface scattering and microtopographic features. *Opt. Eng.* **18**(2) (1979)
4. M. Longuet-Higgins, Statistical properties of an isotropic random surface. *Phil. Trans. A* **250** (1957)
5. J. Stover, *Optical Scattering*, 2nd edn. (SPIE Optical Engineering Press, Washington, 1995)
6. C. Bohren, D. Huffman, *Absorption and Scattering of Light by Small Particles* (Wiley, 1998)
7. T. Draine, P. Flatau, Discrete-dipole approximation for scattering calculations. *J. Opt. Soc. Am. A* **11**(4) (1994)
8. S. Sze, *Physics of Semiconductor Devices*, 2nd edn. (Wiley, 1981)
9. P. Horowitz, W. Hill, *The Art of Electronics*, 2nd edn. (Cambridge University Press, 1989)
10. S. Soares, *Nanoscale Non-Contact Laser Measurement of Precision Machine Tooling and Optical Surfaces, Photooptics 2018* (Madeira, Portugal, 2018)

Chapter 2

Centimeter-Resolution Long-Distance Optical Fiber Monitoring



Felipe Calliari, Luis Herrera, Jean Pierre von der Weid
and Gustavo Amaral

Abstract The worldwide dense deployment of optical fiber networks has been driven by the desire of higher transmission capacity necessary for high-level data communications. Supporting all the higher Open Systems Interconnection (OSI) layers is the physical layer represented by the optical fiber, which, unfortunately, suffers from mechanical fragility. The robust operation of the whole network, therefore, can be jeopardized by mundane events such as the strangling of the fiber cable caused by a passing truck on a highway or by a break due to the action of rodents. In order to deal with this issue, physical layer supervision is of the utmost importance and, over the years, reflectometry techniques have been developed and upgraded so that the protection of the optical fibers is ensured. This chapter performs a brief revision of such reflectometry techniques with special focus on the Optical Time Domain Reflectometry and on one of its most interesting recent developments, the Photon-Counting Optical Time Domain Reflectometry. It further presents a technique for centimeter-resolution long-distance measurements of optical fibers in practical times and discusses its performance in detail.

2.1 Introduction

Although fiber optics were suggested as the best choice for data transmission in 1966 [1], only after 1980s optical communications systems became commercially viable [2]. This was only possible due to the reduction of optical fiber losses as well as the development of semiconductor lasers and optical amplifiers. Undoubtedly, the major

F. Calliari (✉) · L. Herrera · J. P. von der Weid · G. Amaral
PUC-Rio, Rua Marquês de São Vicente, 225, Rio de Janeiro, Brazil
e-mail: felipe.calliari@opto.cetuc.puc-rio.br

L. Herrera
e-mail: luisyno@opto.cetuc.puc-rio.br

J. P. von der Weid
e-mail: vdweid@opto.cetuc.puc-rio.br

G. Amaral
e-mail: gustavo@opto.cetuc.puc-rio.br

© Springer Nature Switzerland AG 2019
P. Ribeiro and M. Raposo (eds.), *Optics, Photonics and Laser Technology 2018*, Springer Series in Optical Sciences 223,
https://doi.org/10.1007/978-3-030-30113-2_2

factor in the advent of the “information age” was the development of fiber optics along with microelectronics technology.

The advantages of optical fibers over other transmission methods (copper cables, satellites, etc.) are quite relevant: immunity to electrical or magnetic interference, low weight in relation to metallic cables, low manufacturing cost and extremely broad transmission bandwidth, just to name a few. Although its great reliability, optical fibers can sometimes be damaged, by a variety of causes [3, 4]: ships can break optical fibers or marine animals can chew the fiber protective coating in the ocean, for example; on land, optical fiber link follow physical infrastructures such as highways, railroads, and electric power transmission lines which are exposed to unexpected pitfalls due to works, storms, or accidents that can end up damaging the fiber cable.

In long-distance optical telecommunications links, fiber monitoring is crucial due to the negative impact that faults in the optical fiber have on the power budget of the links, which substantially limits the achievable transmission rates. The impact is not limited, however, to high data rate applications, but extend to essential services such as banking, telephone, and internet service links, which can be interrupted.

Amongst the plethora of techniques that can be used for determining the occurrence of a fault as well as locating it along the fiber link, reflectometry systems are an extremely attractive one. The physical phenomenon underlying such class of techniques is Rayleigh scattering, which causes light from a propagating probe pulse along the fiber to be scattered back to the sender, where it can be detected and processed to yield useful information about the fiber’s conditions.

In this chapter, apart from a comprehensive revision on reflectometry techniques applied to fiber optics, recent advancements in optical fiber monitoring using the so-called photon-counting Optical Time Domain Reflectometer will be presented. The technique offers ground-breaking compromise between dynamic range, spatial resolution, and monitoring times by combining independent photon-counting OTDR strategies and a powerful fault location algorithm. In order to provide the reader with all the background necessary to grasp the concepts involved, a brief revision of the main optical fiber monitoring strategies based on reflectometry will be presented in Sect. 2.2, followed by a detailed review of photon-counting OTDR protocols in Sect. 2.3. In Sect. 2.4, the foundations of the fault location algorithm, as well as its practical implementation, are provided. Section 2.5 closes the chapter by presenting a technique for centimeter-resolution optical fiber monitoring and discussing its results.

2.2 Reflectometry Strategies for Optical Fiber Monitoring

The Rayleigh scattering phenomenon was discovered in 1871 by John William Strutt (Lord Rayleigh) during observations of the solar light scattering in the Earth’s atmosphere [5, 6]. It is associated to the scattering of electromagnetic radiation from particles with radii much smaller than its wavelength [2, 7], and relates the scattered power to the wavelength as

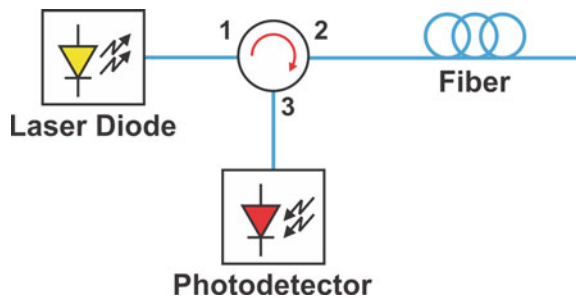
$$R \sim \frac{1}{\lambda^4}. \quad (2.1)$$

In optical fibers, the disordered nature of the silica glass medium causes microscopic fluctuations in the local density and, thus, in the index of refraction experienced by the guided light wave. Since such microscopic fluctuations are still orders of magnitude smaller than the wavelength of the transmitted light (usually in the near infrared range) energy losses due to the Rayleigh scattering of light are observed [2]. Even though the scattering is omnidirectional, a small portion of the scattered light is collected into the acceptance cone of the fiber and propagates in the the opposite direction of the incoming light [8].

Since the process of Rayleigh scattering is predominantly elastic, the wavelength of this back-propagating light is the same as the incoming light, so no spectral filtering can separate them [9]. An optical circulator, on the other hand, is a device that allows separation of light signals that travel in opposite directions inside the fiber, and can be used to evaluate the Rayleigh backscattered portion of light. By putting together a light source, an optical circulator, and a photodetector as in Fig. 2.1, a system for measuring the intensity of Rayleigh backscattered light is created. Furthermore, since losses and reflections along the fiber will be experimented by the backreflected light, it can be used to evaluate the conditions of the fiber while only accessing one of its input ports. Several of such single-ended (as it only access one of the ends of the fiber) fiber probing techniques making use of Rayleigh backscattering have been developed over the years.

The Optical Time Domain Reflectometry. The first single-ended method designed for gaining information of undesirable losses in an optical fiber was the Optical Time Domain Reflectometer (OTDR) [11]. The OTDR can be interpreted as a device that measures the impulse response of the fiber, i.e., its reflectivity as a function of time. To do so, a probe pulse is transmitted and the Rayleigh backscattered power is measured in the photodetector connected to port 3 of the circulator as a function of time, as seen in Fig. 2.1. Also, in order to measure the impulse response of the fiber the condition of one pulse inside a fiber at a time must be satisfied [24]. Given the index of refraction of the fiber and the speed of light in vacuum, one can associate the time to the position of the fiber where the scattering took place and, thus, losses

Fig. 2.1 Simplified block diagram of a general reflectometry technique for optical fiber probing, adapted from [10]



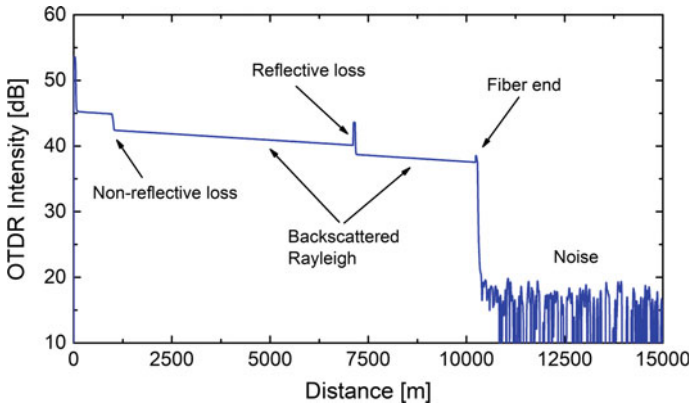


Fig. 2.2 Example of OTDR profile showcasing common events that can be identified. The profile has been acquired with a standard OTDR device Anritsu MT9083C [12]

can be located. Figure 2.2 depicts an example of OTDR profile showcasing typical events that can be identified.

The accuracy of an event location is determined by the temporal width of the propagating pulses: as it approaches zero, the trace produced by the OTDR approaches the impulse response and the accuracy is enhanced. On the other hand, light sources have limited output peak powers and, as the width approaches zero, the total optical power carried by the probe pulse also approaches zero. Due to the intrinsic attenuation of the fiber, only short fiber lengths can be reached if the width of the probe pulse is too short.

Although enlarging the probe pulse allows for longer fibers to be evaluated, the accuracy is compromised. The *spatial resolution*, or accuracy in locating a loss event, and the *dynamic range*, or the maximum amount of optical losses that can be withstood by the probe pulse to reach a position of the fiber, constitute a compromise relationship for OTDR devices [2, 8]. It is interesting to note that, in order to achieve better spatial resolution, the quality of the optoelectronic devices must be enhanced. In other words, in order to be able to generate and detect short optical pulses, both a high-speed modulator and a high-speed detector must be employed.

An experimental and visual evidence of the the practical compromise relationship between the spatial resolution and the dynamic range is presented in the two panes of Fig. 2.3. The results correspond to the OTDR profiles of the same fiber, but with different pulse widths of the probing pulse. While the results of the upper pane correspond to a 200 ns-wide pulse yielding a 24 dB dynamic range, the results of the lower pane were acquired with a 20 ns-wide pulse yielding a reduced 13 dB dynamic range. As the dynamic range decreases from one result to the other, however, it becomes clear from the inset of the figures that the spatial resolution has been greatly enhanced and the position of the reflection has been located with higher precision.

The Optical Frequency Domain Reflectometry. While OTDRs use a pulsed optical source to measure an approximation of the impulse response of the fiber, the

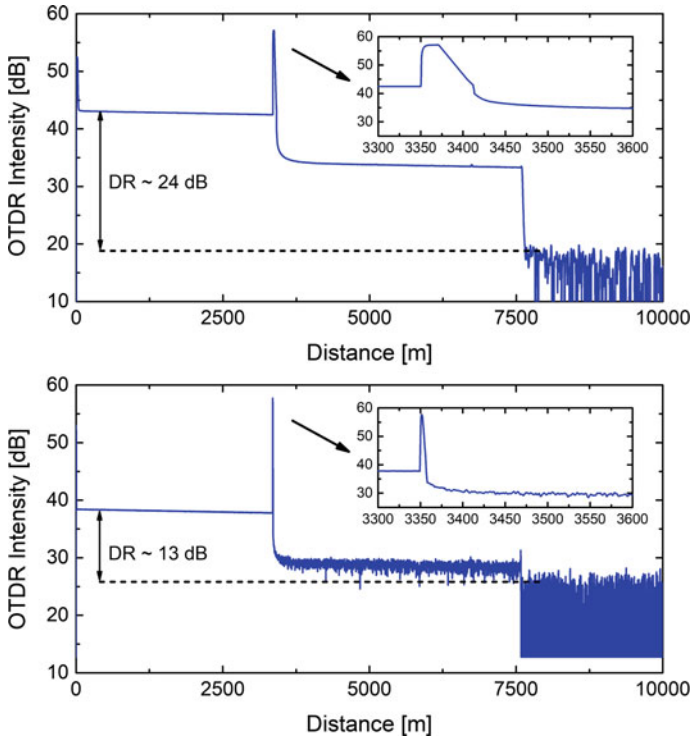


Fig. 2.3 Experimental and visual evidence of the trade-off between spatial resolution and dynamic range. In both panes, the same fiber is probed but with a 200 ns-wide pulse in the upper pane and 20 ns-wide pulse in the lower pane. As the dynamic range decreases from 24 to 13 dB, the accuracy in determining the position of the reflection peak (inset of both panes) increases. The profiles have been acquired with a standard OTDR device Anritsu MT9083C [12]

Optical Frequency Domain Reflectometer (OFDR) uses a continuous-wave (CW) light source to measure the transfer function of the fiber [13–15]. It is a well-known fact from theory of systems that the impulse response and the transfer function are uniquely related by the Fourier Transform, so the positions of losses can be determined by analyzing its response in the frequency domain.

Even though many different implementations of OFDR can be found in the literature, two distinct sets can be identified: coherent (C-) and incoherent OFDR (I-OFDR). In the former, the wavelength of a CW coherent light source is swept and the system measures the optical heterodyne beat between a portion of the probing light and the backscattered light [14]. In the latter, the frequency of a radio-frequency optical sub-carrier is swept and the beat between reference and backscattered signals is performed in the electrical domain instead of in the optical domain as for C-OFDRs [16]. In either case, utilizing the Fourier Transform, the frequency profile can be transformed back to the time domain so the events can be located in the fiber.

Also, the speed of the wavelength or sub-carrier frequency sweep defines the spatial resolution of OFDR techniques.

A third flavor of OFDR is the Step-Wise Incoherent OFDR, where instead of using the sweeping speed of the optical sub-carrier to generate the heterodyne beat between reference and measured signals, the steady-state response of the fiber for each frequency value is measured [17]. It has been recently shown that this method allows for characterization of optical fibers even when the frequency range of sweep is not sufficient for a Fourier Transform-based translation from frequency to time. In such case, the analysis of the complex spatial-dependent phasor output from the measuring apparatus is analyzed in the frequency domain with a high-dimensional signal processing operator to yield relevant spatial resolution.

Modifications to the Original OTDR. Several modifications have been proposed and demonstrated using the OTDR as a core idea, with the objective of increasing the dynamic range and signal-to-noise ratio (SNR), while enhancing the spatial resolution. The coded OTDR (COTDR) is a proposition that overcomes the tradeoff between SNR and resolution [18, 19]: instead of a single pulse, random code sequences are used to modulate the probing light, and the reflection point is located by correlating the backscattered light with the delayed version of the code.

Due to the difficulty in generating the codes necessary for COTDR, the Chaos OTDR was developed, where the random codes are generated by perturbing a semiconductor laser with optical injection [20, 21]. These methods, however, consider light sources with large spectral bandwidths and might not be fit for monitoring wavelength-division-multiplexing (WDM) networks. The Tunable-OTDR, on the other hand, makes use of highly coherent optical sources and optical amplifiers to generate spectrally narrow pulses adapted to WDM networks [22].

In the classical OTDR technique, detection involves the use of p-i-n and Avalanche Photodiodes since the backscattered optical power is continuous at the detector. The maximal Spatial Resolution, i.e., the minimum distance detectable between two events, therefore, is dependent on the detector's bandwidth and on the pulse's temporal width. Unfortunately, the reduction of the bandwidth has a direct negative impact on the noise-equivalent power of the photodetector and imposes a limitation NEP and imposes a limitation on the measurement range.

An alternative to increased Spatial Resolution is the Photon-Counting OTDR, which employs single photon detectors. Such devices offer better sensitivity since the NEP is minimal given that a single photon is capable of generating a detection current P. Eraerds et al. [23]. With the advent of the Geiger-mode Single Photon Detector (SPD) in the telecommunication wavelength, the Photon-Counting OTDR was proposed. Operating in the gated mode, such SPDs can attain high photon detection efficiencies and extremely low dark count rates [23]. The gated operation requires, however, an intelligent management system to reduce monitoring periods and enhance the acquisition of statistically relevant data [24].

The principal advantage of the Photon-Counting OTDR over standard OTDR strategies is an alleviated compromise relationship between spatial resolution and dynamic range. The core to this enhancement is the high sensitivity of the single-photon detector, which allows for the detection of backscattered light from short

pulses even after long distances [23]. Slower data acquisition is, however, a drawback of the Photon-Counting OTDR and also is founded on the characteristics of the single-photon detector, which undergoes a so-called dead time after a detection during which no data is acquired [25, 26].

To try and minimize the low data acquisition rate of Photon-Counting OTDRs, [24] used a train of gates that triggers the single-photon detector multiple times; the results show decent achievable dynamic range with minute time measurements. On a parallel work, [27] focused on increasing the resolution of Photon-Counting OTDR measurements, reaching the mark of centimeter resolutions. These two works not only highlight the results of Photon-Counting OTDR techniques but are also of paramount importance to the results presented in Sects. 2.4 and 2.5. Therefore, they will be revised in the following.

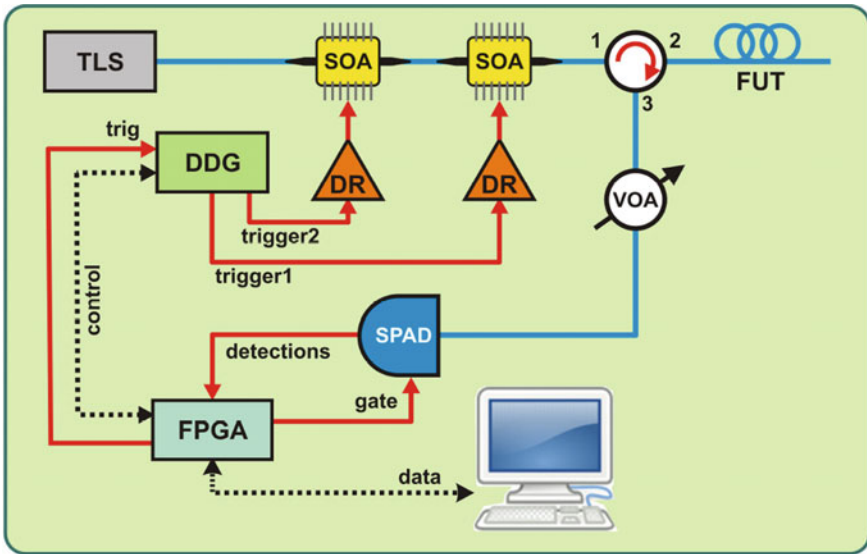
High Dynamic Photon-Counting OTDR. The High-Dynamic Photon-Counting OTDR refers to the technique presented in [24], which makes use of semiconductor optical amplifiers (SOA) to carve high peak power pulses in a continuous wave signal output from a highly coherent tunable laser source (TLS). Given the input of the TLS, the stimulated emission process will cause the amplified light to be concentrated within the original TLS bandwidth even if the SOA is saturated [28]. Therefore, to achieve maximum output peak power, an electronic driver (DR) capable of supplying 50 ns-wide current pulses of 2 A drives the SOA.

By synchronizing the pulse emission of the SOA with a train of enabling pulses directed to the detector, an FPGA board time-tags the detections which are processed by a personal computer and create the fiber profile. The experimental setup of the High-Dynamic Photon-Counting OTDR is presented in Fig. 2.4, where a variable optical attenuator can be seen before the single-photon detector input; its role is to prevent the non-linear operation of the device by controlling the average mean number of photons that impinge on the detector.

The time-tagging of the detection pulses performed by the FPGA board is further detailed in Fig. 2.5. Since the detector must undergo a dead time period after detections, the train of enabling pulses, or gates, must abide by this time. In order to acquire detections from all the positions of the fiber, the optical pulse and the train of gates is shifted in time; this is performed with the help of a digital delay generator (DDG) that creates pulses with an arbitrary time delay with respect to an input trigger. Furthermore, each detection is associated with a 16 bit word composed by the Gate and Pulse number, respectively, which are processed by a personal computer.

The results achieved by the High-Dynamic Photon-Counting OTDR are considerable in terms of dynamic range, resolution, and data acquisition rate. The system allows for minute-long measurements that output results with approximately 5 m resolution and a 32 dB dynamic range. Even if the spatial resolution is not in the centimeter range, With all things considered, this technique figures as a good candidate for quickly evaluating an optical fiber.

High-Resolution Photon-Counting OTDR. A Time-to-Digital Converter (TDC) is a device that allows measurements of extremely short time spans between two events and can be used to increase the time resolution of a detection event. While for the High-Dynamic Photon Counting OTDR, detections are associated to the time stamp



TLS: Tunable Laser Source, **SOA:** Semiconductor Optical Amplifier, **D:** Driver, **DDG:** Digital Delay Generator, **SPAD:** Single Photon Avalanche Detector.

Fig. 2.4 Experimental setup for the high-dynamic photon-counting OTDR, adapted from [29]

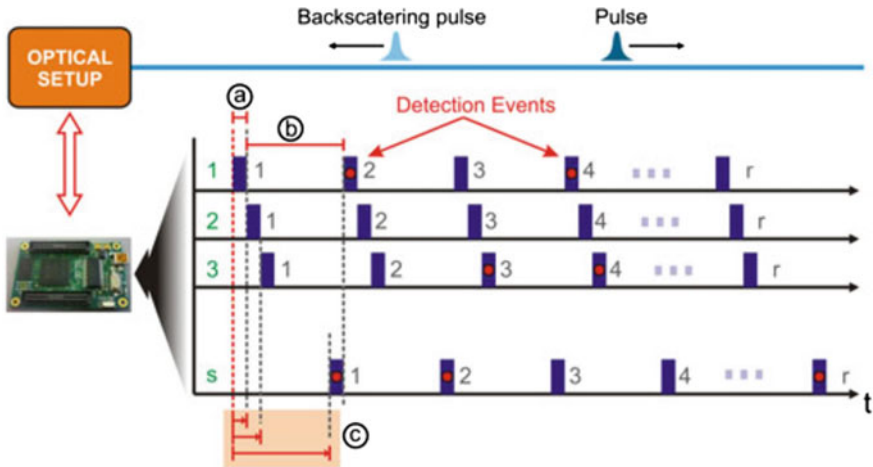


Fig. 2.5 Protocol for detection indexing, or time-tagging, employed by the FPGA board [24]. The parameters a , b and c represent the gate window, the dead time and the maximum delay between pulses, respectively, adapted from [24]

of the gate pulse that they correspond to, in the High-Resolution Photon-Counting OTDR a TDC is responsible for time-tagging the detection with respect to the optical pulse. Furthermore, the optical probes generated by the light source in this setup are

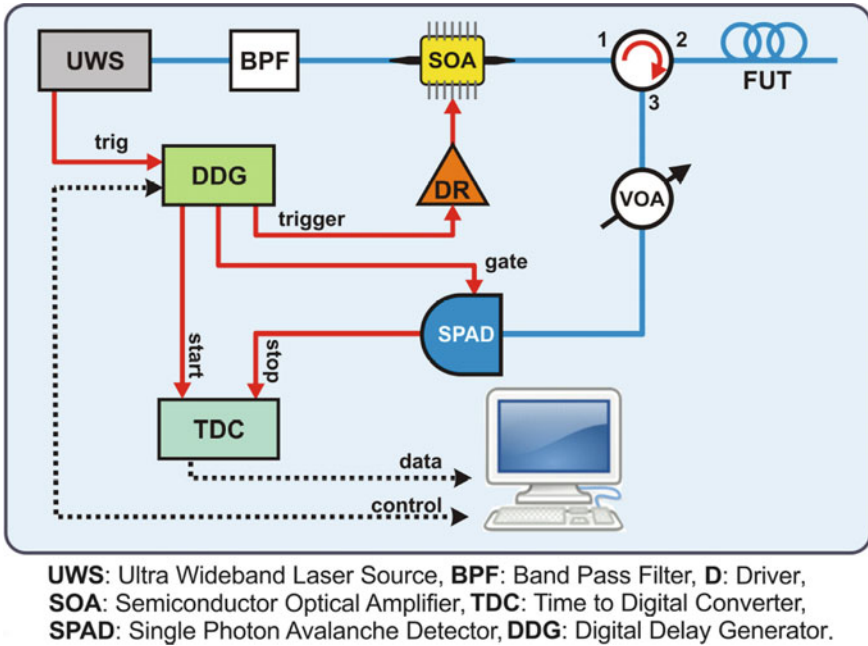


Fig. 2.6 High-resolution photon-counting OTDR experimental setup depicted as a simplified block diagram, adapted from [29]

extremely short, in the order of hundreds of femtoseconds. The conjunction between short pulse generation, high single-photon detector sensitivity, and fine TDC time-tagging allows the High-Resolution Photon-Counting OTDR to achieve centimeter resolution even in long-distance measurements.

The simplified block diagram of the experimental setup depicted in Fig. 2.6 shares many similarities with the one in Fig. 2.4, with the main differences being the replacement of the TLS by an Ultra Wideband Optical Source (UWS) and the time-tagging system, which instead of being conducted by the FPGA board, is a responsibility of the TDC. The transform-limited pulses generated by the UWS are 115 femtoseconds-wide and cover a broad spectral range from 800 to 1600 nm approximately [30]. In order to operate under these conditions, the pulse repetition rate is internally set as 5.8 MHz, which may conflict with the restriction of one pulse inside the fiber at a time. In order to reduce the pulse repetition rate and satisfy this condition, an SOA is connected after the UWS.

The SOA is triggered by an electronic driver (DR) that generates 4 ns-wide 600 mA pulses with the goal of allowing the desired optical pulse to be transmitted instead of actually amplifying it. Again, the synchronization between the pulses generated by the UWS and the pulses that trigger the SOA is performed by a Digital Delay Generator (DDG). A personal computer is responsible for interpreting the detections

of the TDC and composing the fiber profile, which can achieve a spatial resolution of 2.8 cm.

Unfortunately, the dynamic range achieved with this setup is limited by 14 dB due to the spectral filtering performed by an optical band-pass filter included in the experimental setup between the UWS and the SOA. The role of this filter is to shorten the bandwidth of the optical pulses so that tunability can be enforced [10]. Additionally, the detection mechanism, even though responsible for the high resolution achieved, only allows a single detection to be acquired per optical pulse transmitted through the fiber. This, in turn, reflects in a low rate of data acquisition.

2.3 Fault Location Algorithms

The main objective of a physical layer supervision technique such as the OTDR is to identify the potential faults that cause undesirable power losses over the transmission channel. From the set of all possible events that can be identified in an OTDR, it is possible to differentiate between three subsets: the events which are accounted for and are necessary for the link operation—this includes power splitters, connectors, filters, etc; undesirable events that cause minor power losses—this subset includes losses due to bending of the fiber where the radius of curvature is close, but not smaller, to the critical radius; and, finally, the events which are unaccounted for and substantially affect the power budget of the optical link—fiber breaks, extreme bends, and defective connectors are included in this subset.

Given a fiber profile, measured with the supervision technique that best suits the optical network, it is the role of an operator to distinguish between these three sets. This is extremely relevant since, if events that fall into the third set are identified, an in-field repairing unit must be scheduled so that the link operation is not compromised either for too long or not at all: recent developments on Software Defined Optical Networks (SDON) allow re-routing of data traffic to avoid defective fibers while maintaining quality of service (QoS) [31].

It is also relevant, in this context, that the information regarding the position of the fiber to be repaired is accurate. As detailed in previous sections, this is intimately tied to the available spatial resolution of the supervision technique. However, it is also related to considerable work force spent by the operator, especially when one considers the dense deployment of optical fibers that the world has experienced. Therefore, a digital signal processing tool capable of identifying the positions of faults in the digitized fiber profile could alleviate the workload and, eventually, hasten and simplify the scheduling process of the in-filed repairing units.

Specifically in OTDR profiles, the power losses are directly associated to abrupt breaks in the trend of the data stream. Thus, in order to automatically identify the faults, an algorithm must be able to identify the positions of such trend breaks and their magnitudes. Furthermore, it is imperative that the algorithm operates even in the presence of noise, since the imperfection of the measurement apparatus will inherently produce noisy fiber profiles. Identification of trend breaks, also known

as step filtering, can be found in the literature, with a handful of algorithms fit to deal with problem [32–36]. Recently, however, a comparative study was conducted where several of such algorithms were tested and the Adaptive ℓ_1 Filter was elected as the most prominent candidate for application in identification of multiple faults in fiber profiles in the presence of noise [37].

In order to analyze the OTDR profile and identify the faults, a simplified model of the expected trend breaks is created, in which the remaining features such as noise contributions from CRN [38], noise contributions from measurement uncertainty [24], or even reflection peaks are considered as noise. The slope of the fiber, on the other hand, is a feature that cannot be neglected, since the algorithm would eventually compensate the continuous power loss with spurious trend breaks and overload the result with false positives. Fortunately, the inclusion of the slope in the model is a simple task and does not compromise the processing [37, 39]. The simplified model of an arbitrary OTDR profile $y(d)$ (written in the logarithmic scale) is then written as:

$$y(d) = \alpha \sum_{i \in F} a_i u_1(d - d_i), \quad (2.2)$$

where α corresponds to the negative slope of the profile (the intrinsic optical attenuation of the fiber), and $d_i \in F$ are the eventual fault positions with amplitudes a_i that belong to the set F of all faults. u_1 denotes the Heaviside function that represents discontinuous steps in the data series.

Identifying the set F is the ultimate goal of the fault location algorithm as it will characterize the fiber in terms of the potential fault events that are present. Since the model is defined and the functions that are expected to compose the profile $y(d)$ are known, the Adaptive ℓ_1 Filter implements a basis pursuit problem that recovers the set F given a larger set \mathbb{F} of all the possible faults. The procedure involves a mixed ℓ_1/ℓ_2 norms minimization, where an equilibrium between the best approximation of the original signal and the least amount of components in the set F is demanded [37]. The result is a highly accurate identification of not only the faults positions, but also its magnitudes; Fig. 2.7 presents the results of the Adaptive ℓ_1 Filter algorithm on a standard OTDR trace, where the selections accurately match the positions and magnitudes of the faults.

The versatility provided by such a signal processing tool is extremely broad from the perspective of optical network supervision. First and foremost, it alleviates the workload of an operator, which may access only the output of the filter as an event list instead of inspecting the profile in search for the positions and magnitudes of faults. Furthermore, the massive deployment of optical fibers fomented by the ever-increasing need for higher data transmission capacity poses a problem when one considers supervising and managing all the links; an automated procedure can process multiple profiles and identify the ones which require dedicated attention. In the next section, a procedure that allows minute-time long-distance supervision with increased resolution is presented, having the fault detection algorithm at its core.

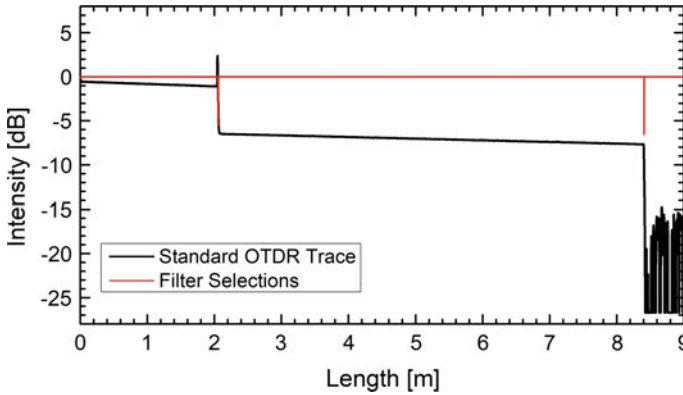


Fig. 2.7 Adaptive ℓ_1 Filter selections for a standard OTDR profile. Faults at 2 and 8.4km have been successfully identified as well as the corresponding magnitudes in dB, adapted from [40]

2.4 Long-Distance Centimeter-Resolution Monitoring Mediated by Fault Location Algorithm

Characterizing an optical fiber with high precision is, as discussed in the previous sections, an important ingredient of modern optical networks, which allows for robust managing of the optical network from the operator's point of view and with higher quality of service to the user. More often than not, it also translates in cost reduction due to the simplification of tasks. In this context, the more precise is the identification and characterization of an event, the better, a desire that often encounters the compromise relation between dynamic range and spatial resolution.

The Photon-Counting OTDR strategies presented in Sect. 2.2 offer distinct but complementing characteristics to this compromise relation, i.e., the High-Dynamic Photon-Counting OTDR offers high-dynamic range (32 dB) while delivering descent spatial resolution (6m), and the High-Resolution Photon-Counting OTDR offers extremely high resolution (3 cm) while exhibiting considerable dynamic range (14 dB). One other feature that distinguishes between these strategies is the speed of data acquisition, which is optimized in the case of the former but extremely slow for the latter. At the same time an algorithm capable of accurately identifying the fault positions and magnitudes was discussed in Sect. 2.3.

By making use of these three elements and concatenating their outputs, it is possible to achieve fast high-resolution measurements of long-distance links. In this hybrid structure, the high dynamic range of the system depicted in Fig. 2.4 and the high resolution provided by the system depicted in Fig. 2.6 are harnessed by mediating their operation with the Adaptive ℓ_1 Filter. A fully automatic monitoring structure was developed [10] and the chart shown in Fig. 2.8 depicts the steps taken by the system to inspect a fiber optical link.

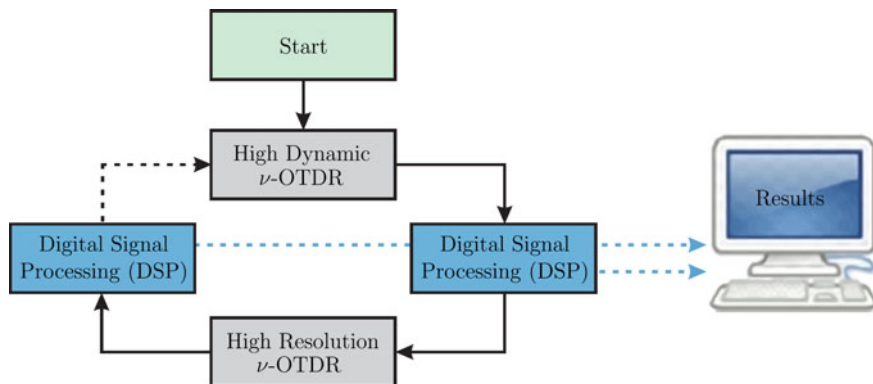


Fig. 2.8 Flowchart of the Hybrid and Automatic ν -OTDR, adapted from [40]

By following the chart depicted in Fig. 2.8, four steps are identified and are as follows. First, the system uses the High-Dynamic Photon-Counting OTDR to obtain a fiber profile with a high acquisition rate and coarse spatial resolution; in a 20 km link, for example, the acquisition rate is still quite high and the system allows the determination of the last point of the profile with a signal-to-noise ratio of 10 dB in just under two minutes [28]. In the second step, the digital signal processing step, the Adaptive ℓ_1 Filter attempts to identify those positions which might contain events; this step is also quite fast with accurate results in just under a minute [24].

In the third step, the High-Resolution Photon-Counting OTDR analyzes the regions around the positions estimated in step two, i.e., those selected by the algorithm; even though this third step would take a substantially longer time than the others if it was to inspect the whole fiber, the fact that it only needs to acquire data from a small set of positions causes it to be performed in practical times, in the order of few minutes. Finally, the fourth step involves analyzing the results of the third step once again with the fault location algorithm and providing a highly accurate event list to the operator.

Since each of the building blocks of Fig. 2.8 have already been described individually in Sects. 2.2 and 2.3, the setup of Fig. 2.9 presents the full measurement apparatus. It is interesting, however, to note that the individual setups can be easily identified in the full setup.

A personal computer is responsible for managing the coexistence of the apparatuses at each of the measurement steps while also being in charge of processing the acquired data with the fault location algorithm. In the first step, the optical switch's output must be connected to its upper input so that coarse monitoring can be performed. At the same time, the electrical switch outputs a signal to the single-photon detector coming from the FPGA. In the third, step, fine monitoring, both switches modify their outputs triggered by the personal computer [10, 24, 29].

An important comment is regarding the abundance of devices in the setup of Fig. 2.9, where the only device that is shared between both strategies is the

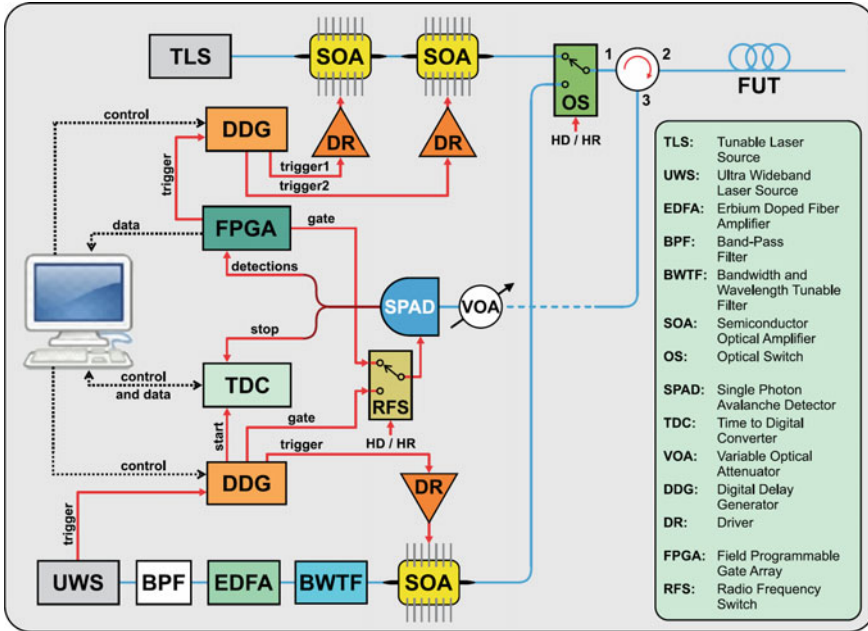


Fig. 2.9 Architecture of the photon-counting OTDR for long-distance centimeter resolution monitoring, adapted from [40]

single-photon detector. Unfortunately, the fine resolution Photon Counting OTDR is extremely dependent on a light source capable of generating short optical pulses. In order to generate such short pulses in time, the spectrum of the optical sources is enlarged, making it iniviable to use the same optical source for both Photon Counting OTDR strategies [41]. That is because the coarse Photon Counting OTDR relies on a narrow linewidth continuous optical source so that the high peak power pulses can be generated. Likewise, the extremely different characteristics of the electrical pulses sent to the semiconductor optical amplifiers on either part of the setup does not allow for the devices to be reused.

2.5 Monitoring Different Optical Fiber Links

In order to demonstrate the capabilities of the long-distance centimeter-resolution system presented in the previous section, different optical fiber links have been selected. These links exhibit characteristics that: allow to validate the possibility of characterizing an event at long-distances with centimeter-resolution in minute-times; allow to observe the possibility of tuning the wavelength of the probing signal so that the system can be adapted to monitor WDM optical networks.

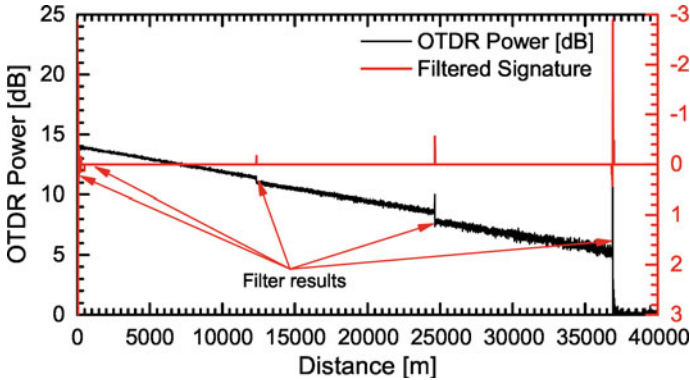


Fig. 2.10 HD- ν -OTDR trace and the filtered signature of the ~ 36 km fiber link, adapted from [40]

First, the monitoring system was tested in a long-haul fiber with total length of 36 km. This fiber is composed of three 12 km fibers connected one to the other, so it is expected that at least three events are identified by the fault location algorithm: 2 events on the connection between the stretches; and 1 event corresponding to the end of the fiber. The acquisition of the High-Dynamic Photon Counting OTDR step takes 150 s. Considering a 6 m spatial resolution achievable by the High-Dynamic Photon Counting OTDR, the 36000 m profile yields a digitized data series containing 6000 points. This profile is processed by the Adaptive ℓ_1 Filter in under two minutes, and the results of both the measured profile and the filter selections are displayed in Fig. 2.10.

It is interesting to note that the filter identified all the expected positions of faults but also outputs a few false positives in the first positions. The first false positive detection of the filter corresponds to the positive step of the beginning of the fiber, i.e., the transition between 0 and the first point measured. Since the filter consistently detects such transitions in all the profiles, it can be neglected for the purpose of fault analysis. The second false positive, however, is not associated to any preconceived features and must not be neglected.

The list of events output from the Adaptive ℓ_1 Filter is the following: 0 m; 1011 ± 6 m; 12303 ± 6 m; 24603 ± 6 m; and 36830 ± 6 m. As already commented, the first positions (0) will be neglected and not used in the analysis. The next step is to transmit the findings of the filter to the High-Resolution Photon Counting OTDR, which will inspect the fiber in the regions around the fault candidates. This step will have two main effects: the first one has already been discussed, and is the centimeter inspection of the events present in the profile; the second, is the eventual elimination of false positives generated by the first run of the algorithm. This is accomplished by analyzing the results once again using the Adaptive ℓ_1 Filter. If there are no faults in this position, the detailed centimeter-measurement will clarify this feature and no selections will be made by the algorithm.

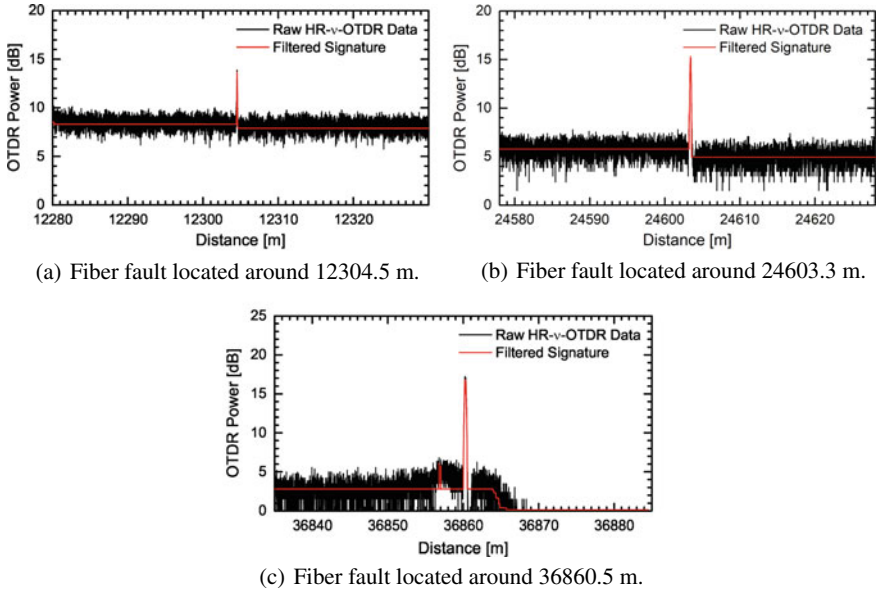
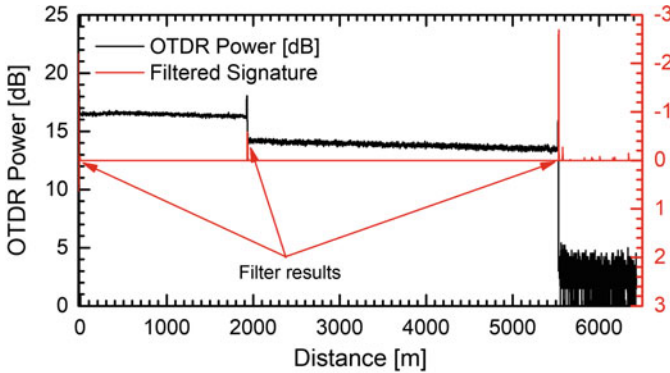


Fig. 2.11 Results of events validated during the final steps of the fault detection routine, adapted from [40]

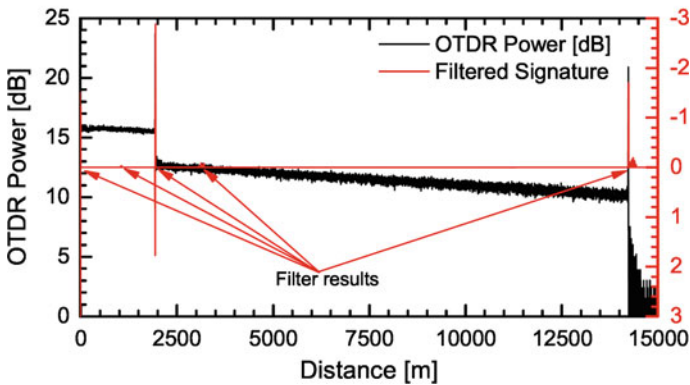
In Fig. 2.11, the results of the third and fourth steps of the chart are presented. Along with the set of positions to be further inspected, the High-Resolution Photon Counting OTDR also requires a range around which the fiber should be probed. This range is arbitrarily set at ± 50 m, which translates into 100 m of fiber per candidate to be analyzed with a 3 cm resolution (around 3000 points in the digitized data series); the total time taken by the third step is 150 s on average for each candidate, while the fourth step lasts around 75 s, also on average, for each candidate. The total accumulated time amounts to 900 s, since four regions were inspected during these steps; the candidate at 1011 ± 6 was dismissed as an event and is not depicted in Fig. 2.11.

In this first test, the fault detection method described in Fig. 2.8 was able to deliver high-resolution event detection in a 36 km fiber under 20 min. Additionally, the event list produced is double-checked by the processing routine and eliminates false positive results from the first algorithm analysis. An important note on this methodology is that, as the number of candidates detected in the second step increases, the total procedure time also increases since all the events must be further investigated in the High-Resolution step.

An important feature of optical network monitoring is adaptability, i.e., being able to evolve with the modern architectures and technologies. Wavelength-division multiplexing is one major recent evolution in optical networking and it is most likely to figure in most of the future deployed optical fiber links. Fortunately, both the High-Dynamic and the High-Resolution Photon Counting OTDR offer what is



(a) HD-v-OTDR trace and the filtered signature of the $\sim 2 + 3.6$ km fiber link at channel 37 of the DWDM grid.



(b) HD-v-OTDR trace and the filtered signature of the $\sim 2 + 12$ km fiber link at channel 40 of the DWDM grid.

Fig. 2.12 Results of events validated during the final steps of the fault detection routine, adapted from [40]

called tunability, the capacity of tune the wavelength of operation to match a specific channel thus adapting to WDM networks.

The evaluation of the monitoring solution in such a network was performed by setting up a simple WDM link with a feeder fiber of ~ 2 km and two wavelength-dedicated fibers of ~ 3.6 and ~ 12 km. These two fibers, in turn, are connected to the feeder fiber through a passive wavelength multiplexer that divides the incoming light spectrum into independent channels. The specific device employed in the experiments is tuned to channels 37 (Fig. 2.12a) and channel 40 (Fig. 2.12b) of the ITUT G.694.1 WDM grid. In Fig. 2.12, the monitoring and fault finding results for the first two steps in the monitoring procedure are depicted. Notable is the absolute distinction between the measurement of channel 37 (Fig. 2.12a) and channel 40 (Fig. 2.12b), indicating that the probing signal is contained within the bandwidth ascribed to the respective channel and does not leak to adjacent channels.

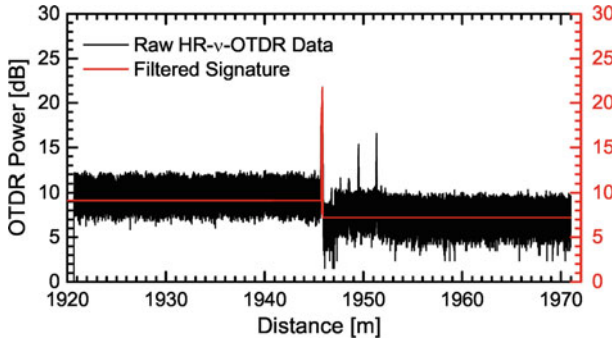
Two distinct features of the fiber profiles displayed in Fig. 2.12 that deserve comments are: the false positives detected by the fault location algorithm after the fiber end, apparent in Fig. 2.12a; and the increased counts right after the reflection peak in the end of the fiber, apparent in Fig. 2.12b. These two effects are intimately related to the intrinsic physical properties of the single-photon detector and can be either bypassed or understood by characterizing and monitoring the operation of this device.

The false positives are associated to the noisy trend after the fiber end that, unfortunately, are interpreted by the algorithm as trend breaks. This feature, however, can be easily neglected in the final analysis since the number of counts after the end of the fiber in the single-photon detector is equal to the expected when considering the detector's dark count rate, a characteristic described in the device's datasheet [42]. Therefore, by monitoring the number of counts, the measurement apparatus is capable of identifying the point after which the counts are not associated to Rayleigh backscattered photons but, rather, merely to the detector's noise. Detections associated to positions associated to dark counts are neglected after the signal processing step and are not fed to the subsequent monitoring steps.

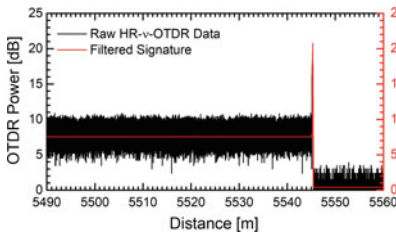
The increase in the number of counts after the high-reflectivity event that creates a exponential skirt at the end of the fiber profile in Fig. 2.12b is associated to the afterpulse effect of single-photon detectors [43]. This effect is related to the trapping of charges inside the semiconductor junction and scales with the input power directed to the detector. Since the appearance of such high-reflectivity events is not controllable by the monitoring apparatus but is associated to the quality of connections along the link, there is no means of bypassing such feature in the profile. Furthermore, it is not unlikely that the fault location algorithm will create false positives due to this feature since it is not described in its simplified model of the profile. Indeed, it can be noted that a set of false positives has been selected by the algorithm. Fortunately, the false positives associated to the afterpulse effect are present in the vicinity of a real event, and can be later processed by the High-Resolution monitoring without expending otherwise unnecessary monitoring time. Additionally, the afterpulse effect has less impact on the High-Resolution measurement due to the method associated to the data acquisition in this step [44].

In Fig. 2.13, the results of the third and fourth step of the monitoring routine are depicted. The false positives that have been identified in the High-Dynamic OTDR profile of channel 40 are not depicted, but have been excluded from the event list after analysis in the fourth step. Also, since the detection of the connection between the feeder fiber and the wavelength multiplexer device around 2 km has been identified in both traces, it is depicted only once in Fig. 2.13a.

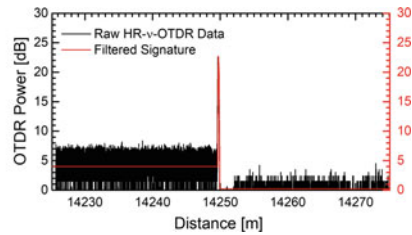
The monitoring time of both channels was performed in approximately 25 min. The fact that two false positives have been identified by the fault location algorithm in the first analysis caused the overall time to increase significantly. Nevertheless, the possibility of tuning the wavelength of operation in both the High-Dynamic and High-Resolution measurement steps allowed for the system to accurately probe the WDM link and identify the correct event positions and magnitudes with high precision.



(a) High-resolution measurement around 1945 ± 6 meters.



(b) High-resolution measurement around 5525 ± 6 meters.



(c) High-resolution measurement around 14250 ± 6 meters.

Fig. 2.13 Results of the third and fourth monitoring steps in the WDM optical link. Tunability of the monitoring apparatus allows for measuring each channel individually, adapted from [40]

2.6 Tunability, Coherence, Chromatic Dispersion and Spatial Resolution

In order to be compatible with WDM networks (as specified in ITU-T G.694.1), a tunable OTDR must not only be capable of selecting the center wavelength of emission but also the spectral width of the optical signal. The current Dense WDM (DWDM) network channels operate with either spectral widths of 0.8 nm and 40 channels, or 0.4 nm and 80 channels, although 0.2 nm standard (the so-called Ultra-Dense WDM) have also been proposed [45]. The necessity of a tunable OTDR for monitoring WDM networks becomes clear by the result of Fig. 2.14a which shows an OTDR profile of a WDM network composed by a feeder fiber, a WDM splitter and four user fibers. Since the source used for this measurement has a broad spectral width, the backscattered power of all channels was overlapped and could not be distinguished in the resulting profile except by the end-fiber reflection peak.

In contrast with Fig. 2.14a, b presents the profiles of the same WDM network but when a spectrally-tailored optical source was employed for the measurements. The result is two-fold: the links can be individually monitoring without overlap between the Rayleigh backscattered power from one channel to an adjacent one;

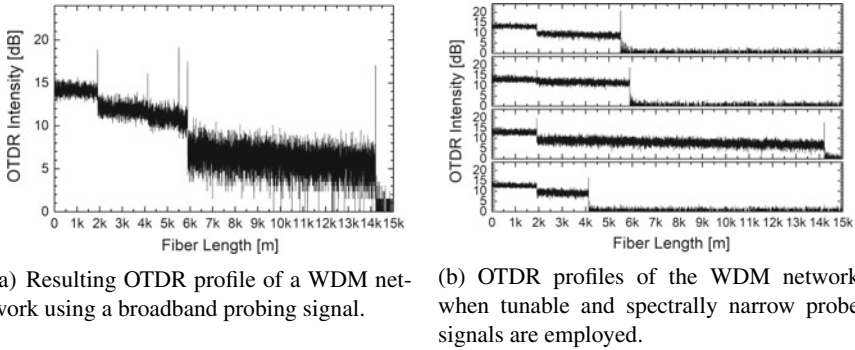


Fig. 2.14 Results of events validated during the final steps of the fault detection routine, adapted from [40]

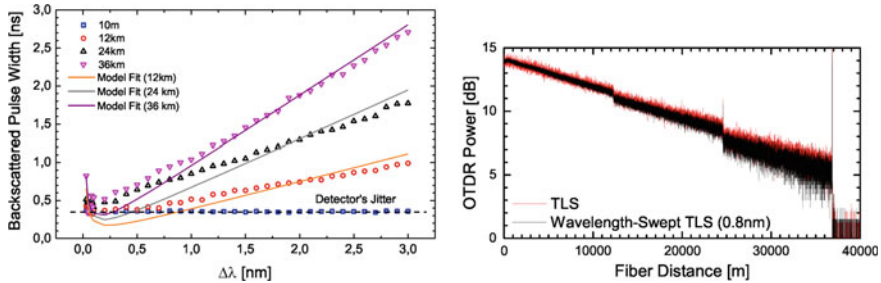
monitoring each link individually causes the monitoring time to increase substantially. As commented in Sect. 2.2, tunability is easily enforced for the High-Dynamic Photon Counting OTDR since it depends solely on the TLS connected to the setup. For the High-Resolution measurements, tunability involves the inclusion of a tunable filter; the high-resolution achievable can thus be compromised by the spectral width of the pulse.

The narrower the spectrum of the optical probing pulse tailored by the tunable filter, the worse the spatial resolution will be; this is because the transform-limited pulses will broaden in time as the spectrum is filtered. This creates a fine balance between three distinct effects of the measurement: using a spectrally broader probing pulse reduces the contribution of Coherent Rayleigh Noise (CRN) on the OTDR profile [37]—CRN are the fluctuations in intensity due to the interference caused by the superposition of several light waves arriving at the detector with random phases [46]—; using a spectrally narrow pulse enlarges the pulse in time reducing the spatial resolution [47]; spectrally broader pulses are more affected by chromatic dispersion while propagating through the fiber, which, in turn, affects the spatial resolution. To illustrate these effects, Fig. 2.15a depicts experimentally measured spatial resolutions for different spectral widths tailored by the tunable filter. At the same time, the CRN effect on an OTDR profile measured with the High Dynamic Photon Counting with narrow and widened spectrum is depicted in Fig. 2.15b.

In Fig. 2.15a, the detector’s jitter is indicated as a reference, since it may also limit the spatial resolution. Additionally, the model used to fit the experimental data was the following [40]:

$$W_p = \Delta\lambda \cdot a + \frac{b}{\Delta\lambda}, \quad (2.3)$$

where a corresponds to $2 \cdot L \cdot D$ and b corresponds to the pulse width of the Fine Resolution Photon Counting OTDR UWS at full bandwidth. Also, we have used an approximation for the dispersion factor D and assumed that it remained constant



(a) Achievable spatial resolution as a function of linewidth for different fiber lengths.

(b) CRN impact on OTDR profile. 600 kHz linewidth laser (red) and linewidth-swept laser [22] (black).

Fig. 2.15 Impacts of spectral tailoring the probing pulse on the achievable spatial resolution and measurement noise, adapted from [40]

within $\Delta\lambda$. From the results, it is clear that the slope for $\Delta\lambda \geq 1$ nm is well fitted, but the experimental results below this value show some inconsistencies with the fit. We conjecture that this behaviour may arise from the fact that the pulse peak power for reduced $\Delta\lambda$ is very low and the measurement results may be distorted. The back-reflected power also diminishes as the fiber length grows, which is observed in a higher contrast between experimental and fitted data for the longer fibers of 24.6 and 32.8 km. The pulse width enlargement for low values of $\Delta\lambda$ is due to the transform-limited pulse, i.e., it is as short as its spectral bandwidth permits.

2.7 Conclusions

An automatic, highly accurate, and fast optical fiber link monitoring system can alleviate several tasks in optical network managing besides reducing operational expenses. Reflectometry techniques based on Rayleigh scattering have met the need for single-ended probing of optical fibers but further improvements can still be performed to enforce compatibility to different network architectures, allow for coexistence between data and monitoring signals, and increase the speed in which events are evaluated. This chapter provided a brief overview of such techniques, while focusing on the Photon-Counting OTDR, a modification of the original OTDR format where instead of a linear photodetector performing data acquisition, a single-photon detector is employed.

To illustrate the potential of Photon-Counting OTDR strategies and the benefits of a fault finding algorithm, a centimeter-resolution long-distance optical fiber monitoring setup has been presented in detail. Tunability to adapt monitoring to WDM networks is a feature of this monitoring apparatus and the results are achieved in practical times. The device incorporates two Photon-Counting OTDR strategies [24, 27] and a fault finding algorithm capable of accurately identifying fault candidates in

a fiber profile. The employed algorithm enables the automation of the whole process so that event candidates can be evaluated with high resolution; furthermore, it allows for the results to be output automatically without the need of an operator. A video containing an experimental run of the method can be found in [48].

The presented monitoring apparatus has been experimentally verified in long-range, mid-range, and wavelength multiplexed optical fiber links. The process involves four distinct steps:

- the High Dynamic Photon-Counting OTDR quickly acquires the fiber profile;
- the Adaptive ℓ_1 Filter identifies potential candidates in the profile;
- the High Resolution Photon-Counting OTDR evaluates the positions around event candidates;
- the Adaptive ℓ_1 Filter is once again used to determine real candidates and their precise positions.

This work paves the way for low-cost, highly reliable, automatic, and fast monitoring of optical fiber links. Even though the results validate the technique, one important improvement to the current setup would allow for a simplified implementation and reduced total cost. The UWS of the High Resolution Photon-Counting OTDR can be replaced by a gain-switched laser for short pulse generation.

Acknowledgements We would like to thank to the brazilian agency CNPq for the financial support, without which this work would not have been realized.

References

1. K. Kao, G.A. Hockham, Dielectric-fibre surface waveguides for optical frequencies, in *Proceedings of the Institution of Electrical Engineers*, vol. 113 (IET 1966), pp. 1151–1158
2. G.P. Agrawal, *Fiber-Optic Communication Systems*. Wiley (1997)
3. J.P. Kilmner, A. DeVito, H.H. Yuce, C.J. Wiczorek, J.P. Varachi, W.T. Anderson, Optical cable reliability: lessons learned from post-mortem analyses, in *Fiber Optics Reliability: Benign and Adverse Environments IV*, International Society for Optics and Photonics, vol. 1366, (1991), pp. 85–92
4. S. Zemon, A. Budman, T. Wei, E. Eichen, K. Ma, Decay of transmitted light during fiber breaks-implications for break location. *J. Light. Technol.* **12**, 1532–1535 (1994)
5. J.W. Strutt, Xv. on the light from the sky, its polarization and colour. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **41**, 107–120 (1871)
6. J.W. Strutt, Xxxiv. on the transmission of light through an atmosphere containing small particles in suspension, and on the origin of the blue of the sky. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **47**, 375–384 (1899)
7. J.W. Strutt, Lviii. on the scattering of light by small particles. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **41**, 447–454 (1871)
8. D. Derickson, *Fiber Optic - Test and Measurement* (Prentice Hall, 1998)
9. L. Kissel, R. Pratt, Rayleigh scattering elastic photon scattering by bound electrons. In: *Atomic inner-shell physics. Physics of Atoms and Molecules*. 1 edn. (Springer, Berlin, 1985), pp. 465–532
10. F. Calliari, Automatic high-dynamic and high-resolution photon counting OTDR for optical fiber network monitoring. Master's thesis, PUC-Rio (2017)

11. M.K. Barnoski, M.D. Rourke, S.M. Jensen, R.T. Melville, Optical time domain reflectometer. *Appl. Opt.* **16**, 2375–2379 (1977)
12. Anritsu: MT9083 Series - MT9083A/B/C ACCESS master. Technical report, Specifications Sheet (2011)
13. R.L. Jungerman, D.W. Dolfi, Frequency domain optical network analysis using integrated optics. *IEEE J. Quantum Electron.* **27**, 580–587 (1991)
14. J.P. von der Weid, R. Passy, G. Mussi, N. Gisin, On the characterization of optical fiber network components with optical frequency domain reflectometry. *J. Light. Technol.* **15**, 1131–1141 (1997)
15. K. Yuksel, M. Wuilpart, V. Moeyaert, P. Mégret, Optical frequency domain reflectometry: a review, in *ICTON'09. 11th International Conference on Transparent Optical Networks*, 2009. (IEEE, 2009), pp. 1–5
16. G.C. Amaral, A. Baldivieso, J.D. Garcia, D.C. Villafani, R.G. Leibel, L.E.Y. Herrera, P.J. Urban, J.P. von der Weid, A low-frequency tone sweep method for in-service fault location in subcarrier multiplexed optical fiber networks. *J. Light. Technol.* **35**, 2017–2025 (2017)
17. J. Nakayama, K. Iizuka, J. Nielsen, Optical fiber fault locator by the step frequency method. *Appl. Opt.* **26**, 440–443 (1987)
18. N. Park, J. Lee, J. Park, J.G. Shim, H. Yoon, J.H., Kim, K. Kim, J.O. Byun, G. Bolognini, D. Lee et al., Coded optical time domain reflectometry: principle and applications, in *Asia-Pacific Optical Communications, International Society for Optics and Photonics* (2007), p. 678129
19. R. Liao, M. Tang, C. Zhao, H. Wu, S. Fu, D. Liu, P.P. Shum, Harnessing oversampling in correlation-coded otdr (2017). [arXiv:1705.05241](https://arxiv.org/abs/1705.05241)
20. Z. Xie, L. Xia, Y. Wang, C. Yang, C. Cheng, D. Liu, Fiber fault detection with high accuracy using chaotic signal from an soa ring reflectometry. *IEEE Photonics Technol. Lett.* **25**, 709–712 (2013)
21. X. Dong, A. Wang, J. Zhang, H. Han, T. Zhao, X. Liu, Y. Wang, Combined attenuation and high-resolution fault measurements using chaos-OTDR. *IEEE Photonics J.* **7**, 1–6 (2015)
22. D.V. Caballero, J.P. von der Weid, P.J. Urban, Tuneable OTDR measurements for WDM-PON monitoring, in *2013 SBMO/IEEE MTT-S International Microwave Optoelectronics Conference (IMOC)* (2013), pp. 1–5
23. P. Eraerds, M. Legré, J. Zhang, H. Zbinden, N. Gisin, Photon counting OTDR: advantages and limitations. *J. Light. Technol.* **28**, 952–964 (2010)
24. G.C. Amaral, J.D. Garcia, L.E. Herrera, G.P. Temporao, P.J. Urban, J.P. von der Weid, Automatic fault detection in WDM-PON with tunable photon counting OTDR. *J. Light. Technol.* **33**, 5025–5031 (2015)
25. S. Cova, N. Ghioni, A. Lotito, I. Rech, F. Zappa, Evolution and prospects for single-photon avalanche diodes and quenching circuits. *J. Mod. Opt.* **15**, (2004)
26. L.E.Y. Herrera, Reflectometria óptica de alta resolução por contagem de fótons. Ph.D. thesis, PUC-Rio (2015)
27. L. Herrera, G. Amaral, J.P. von der Weid, Ultra-high-resolution tunable PC-OTDR for PON monitoring in avionics, in *Optical Fiber Communications Conference and Exhibition (OFC)*, 2015 (IEEE, 2015), pp. 1–3
28. G.C. do Amaral, FPGA applications on single photon detection systems. Master's thesis, PUC-Rio (2014)
29. L.E. Herrera, F. Calliari, J.D. Garcia, G.C. do Amaral, J.P. von der Weid, High resolution automatic fault detection in a fiber optic link via photon counting OTDR, in *Optical Fiber Communication Conference, Optical Society of America* (2016) M3F.4
30. Santec Corporation, Ultra-Wideband Source UWS-1000H Technical report, Specifications Sheet (2016)
31. F.A. Ghonaim, T.E. Darcie, S. Ganti, Impact of sdn on optical router bypass. *IEEE/OSA J. Opt. Commun. Netw.* **10**, 332–343 (2018)
32. R.L. Brown, J. Durbin, J.M. Evans, Techniques for testing the constancy of regression relationships over time. *J. R. Stat. Soc. Ser. B (Methodological)* 149–192 (1975)

33. W.S. Rea, M. Reale, C. Cappelli, J.A. Brown, Identification of changes in mean with regression trees: an application to market research. *Econ. Rev.* **29**, 754–777 (2010)
34. M. Storath, A. Weinmann, L. Demaret, Jump-sparse and sparse recovery using potts functionals. *IEEE Trans. Signal Process.* **62**, 3654–3666 (2014)
35. E.J. Candes, M.B. Wakin, S.P. Boyd, Enhancing sparsity by reweighted ℓ_1 minimization. *J. Fourier Anal. Appl.* **14**, 877–905 (2008)
36. S.J. Kim, K. Koh, S. Boyd, D. Gorinevsky, ℓ_1 trend filtering. *SIAM Rev* **51**, 339–360 (2009)
37. J.P. von der Weid, M.H. Souto, J.D. Garcia, G.C. Amaral, Adaptive filter for automatic identification of multiple faults in a noisy OTDR profile. *J. Light. Technol.* **34**, 3418–3424 (2016)
38. K. De Souza, Significance of coherent rayleigh noise in fibre-optic distributed temperature sensing based on spontaneous brillouin scattering. *Meas. Sci. Technol.* **17**, 1065 (2006)
39. M. Souto, J.D. Garcia, G.C. Amaral, ℓ_1 adaptive trend filter via fast coordinate descent, in *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2016* (IEEE, 2016), pp. 1–5
40. F. Calliari, L.E. Herrera, J.P. von der Weid, G.C. Amaral, High-dynamic and high-resolution automatic photon counting OTDR for optical fiber network monitoring, in *6th International Conference on Photonics*. Optics and Laser Technology, vol. 1 (PHOTOPTICS, 2018), pp. 82–90
41. P. Anandarajah, R. Maher, Y. Xu, S. Latkowski, J. O’carroll, S. Murdoch, R. Phelan, J. O’Gorman, L. Barry, Generation of coherent multicarrier signals by gain switching of discrete mode lasers. *IEEE Photonics J.* **3** 112–122 (2011)
42. I.D. Quantique, id210 - Advanced System for Single Photon Detection, Technical report, Specifications Sheet, (2011)
43. I. Rech, S. Marangoni, D. Resnati, M. Ghioni, S. Cova, Multipixel single-photon avalanche diode array for parallel photon counting applications. *J. Mod. Opt.* **56**, 326–333 (2009)
44. L. Herrera, G. Amaral, J. von der Weid, Investigation of bend loss in single mode fibers with ultra-high-resolution photon-counting optical time domain reflectometer. *Appl. Opt.* **55**, 1177–1182 (2016)
45. A. Shahpari, R. Ferreira, V. Ribeiro, A. Sousa, S. Ziaie, A. Tavares, Z. Vujicic, F.P. Guiomar, J.D. Reis, A.N. Pinto et al., Coherent ultra dense wavelength division multiplexing passive optical networks. *Opt. Fiber Technol.* **26**, 100–107 (2015)
46. K. Shimizu, T. Horiguchi, Y. Koyamada, Characteristics and reduction of coherent fading noise in rayleigh backscattering measurement for optical fibers and components. *J. Light. Technol.* **10**, 982–987 (1992)
47. A.F. Elrefaie, R.E. Wagner, D. Atlas, D. Daut, Chromatic dispersion limitations in coherent lightwave transmission systems. *J. Light. Technol.* **6**, 704–709 (1988)
48. F. Calliari, G.C. do Amaral, L.E.Y. Herrera, High Dynamic and High Resolution Automatic Photon Counting OTDR (2017), <https://www.youtube.com/watch?v=KQn9Du2l4NQ>. Accessed on 27 April 2018

Chapter 3

Characterization of High Speed Optical Detectors by Using a Mode Separating Fiber



F. J. Achten and D. Molin

Abstract Nowadays the need for high speed data transmission is still growing. In data centers multimode fibers are commonly used, OM4 fibers are in favor because of its large bandwidth. Qualification of these fibers is very important. The optimal choice of measurement components is essential to guarantee proper fiber qualification. This article describes a method to choose the best (commercially) available detector to measure the ‘Effective Modal bandwidth’ (EMB) of OM4 fibers.

3.1 Introduction

Today the fiber operational wavelength is 850 nm (OM4), but ‘wide band multimode fibers’ (OM5), covering the range 850–950 nm are being developed and tested in systems [1–4]. Both OM4 and OM5 fibers are qualified by a detailed standardized measurement method called ‘Differential Mode Delay’ (DMD) [5–7]. The main parameter considered is the Effective Modal Bandwidth (EMB). The EMB is calculated from the ‘un-normalized DMD profile’ [6, 7]. The specification for the minimal EMB value becomes more relaxed when the wavelength increases from 850 to 950 nm. This is due to the fact that the ‘penalty’ caused by Chromatic Dispersion goes down. At 850 nm the EMB specification is 4700 MHz km, at 950 nm EMB is specified at 2450 MHz km.

When a fiber (graded core index, 50 μm core diameter) is optimized at 850 nm, the EMB will be much higher than the specified minimal value (4700 MHz km). However, when the range 850 to 950 nm is considered, the EMB values at 850 and 950 nm will be closer to the specified minimal values. The reason for this is the fact that maximum EMB at a certain fiber graded index core profile is always reached at one particular wavelength. For OM4 fiber this is 850 nm. The optimal wavelength

F. J. Achten (✉)

Prysmian Group, Zwaanstraat 1, 5651 Eindhoven, CA, The Netherlands
e-mail: frank.achten@prysmiangroup.com

D. Molin

Prysmian Group, Parc Des Industries Artois Flandres, 62092 Haisnes, France

© Springer Nature Switzerland AG 2019

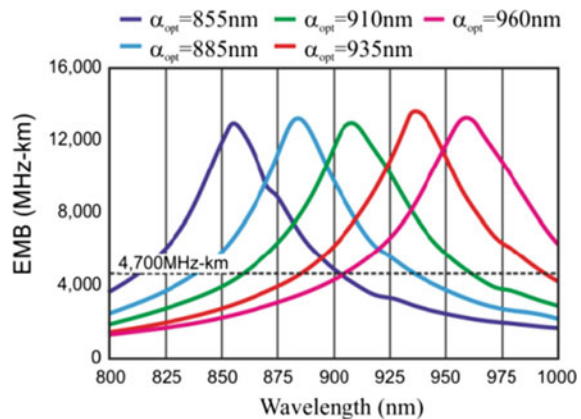
P. Ribeiro and M. Raposo (eds.), *Optics, Photonics and Laser Technology 2018*, Springer Series in Optical Sciences 223,
https://doi.org/10.1007/978-3-030-30113-2_3

can be shifted to -for instance- 900 nm, but the consequence is the EMB reached at 850 nm will go down. OM5 fiber must show high EMB at all wavelengths between 850 and 950 nm. So it makes sense to ‘optimize’ the fiber closer to 900 nm, so both 850 and 950 nm will reach an EMB above the specified value. Figure 3.1 shows EMB versus wavelength for fiber designs with a different ‘Alpha-profile’ (characterized by the parameter α , explained later in this article).

So EMB measurements for OM5 fiber will be much closer to the specified minimal value. This means the precision of the DMD measurement method becomes more relevant. On the launching side (side of the fiber where the laser pulses are launched into), the laser pulse characteristics and the exact position and size of the launch spot are very relevant. On the detector side, two properties are relevant. First of all the detector must be sufficiently fast to detect small changes of the laser pulse shape, after travelling through the fiber. If fibers are measured close to the fiber length used in systems, the laser pulse must be very narrow in time domain (and also in wavelength domain). A typical value for the Full Width Half Max (FWHM) of the laser pulse is 10 ps. A typical value for the detector bandwidth to proper detect this pulse is 25 GHz. The FWQM (Full Width Quarter Max) of the pulse detected by such a detector will be typical 35 ps.

The second property is that the detector must detect all modes that leave the fiber under test. The output of the detector must be a true representation of the spatial and angular power distribution of the pulses leaving the tested fiber. High speed detectors often have a sensitive area smaller than the core of the fiber under test (50 μm diameter). The speed of the detector significantly reduces when the diameter of the sensitive area increases [8]. To detect all light power that is guided by the fiber under test, the internal fiber pigtail of the detector may be ‘tapered’ or ‘lensed’. Purpose is to ‘push’ the light power onto the smaller detector area. The consequence of such a structure is fiber guided light will *not* reach the sensitive area of the detector. This causes ‘selective mode detection’. Another cause of light power loss inside the detector housing is bending of the internal fiber pigtail. This will cause macrobend

Fig. 3.1 Typical dependency of EMB versus wavelength. The specification for EMB is 4700 MHz km. As can be seen from the figure; optimizing at 885 nm results in EMB below specification at 950 nm and optimizing at 910 nm results in EMB below specification at 850 nm [12]



induced losses. To our knowledge, no commercial available detector exists that uses a ‘bend insensitive’ fiber design as internal fiber pigtail. Other causes of selective mode detection may be imperfections of the sensitive area and (slight) misalignments while locking the tapered or lensed fiber pigtail to the sensitive area. If the detector ‘suffers’ from selective mode detection, the measured DMD profile of the OM4 fiber under test is not fundamentally correct, and as a consequence the computed EMB value is less accurate.

This article describes a method to characterize detector performance, so to judge whether the detector can be used to qualify OM4 fiber [9]. OM5 fiber is not considered, as some of the available detectors are not sensitive to 950 nm. The method is tested on five commercially available detectors from different manufacturers. These detectors are all specified for multimode fiber, input is a FC-PC connector. Specified with a very high optical bandwidth of at least 10 GHz. A very high bandwidth is mandatory to qualify short length of OM4 fiber, for instance close to the maximum system length of 500 m. All detectors are fiber pigtailed with an internal 50 or 62.5 μm core Graded Index fiber. Also a low bandwidth detector is used as a reference. This detector cannot detect the individual pulses (too slow), but can measure the power distribution on a near DC level. For now the detectors are investigated at 850 nm, but the same method (for the 950 nm sensitive detectors) can be used for other wavelengths in the range 850–950 nm.

3.2 Characterization Method

First the standardized measurement method called ‘Differential Mode Delay’ (DMD) is explained. This measurement is tightly standardized by *TIA* [6] and *IEC* [7], especially the laser requirements.

The basic idea is to launch laser pulses into the core region of the OM4 fiber. The way these pulses propagate through the fiber core depends on the optical route the pulses will follow through the core. If the core index profile is theoretically correct, all possible routes will end up in the same ‘time of flight’ for the pulses to reach the other end of the fiber. So if the energy of the launched pulse travels through more routes, all energy will leave the fiber at exactly the same moment in time. The relative shape of the pulse in time domain will not change through the fiber (of course the pulse is attenuated). In this case the bandwidth (EMB) of the fiber is infinite. In practice however... a launched laser pulse will always see more than one optional route, and these routes will cause a slightly different time of flight. In Fig. 3.2 this situation is shown schematically.

The maximum number of ‘routes’, for the type of fiber considered at 850 nm is 18, as maximum 18 mode groups can exist. This will be discussed later in this article.

The eventual observed EMB of the fiber depends not only on the fiber itself. It also depends on the characteristics of the launched laser pulse and the characteristics of the detector. Both must be known very precise. Only if the characteristics of both

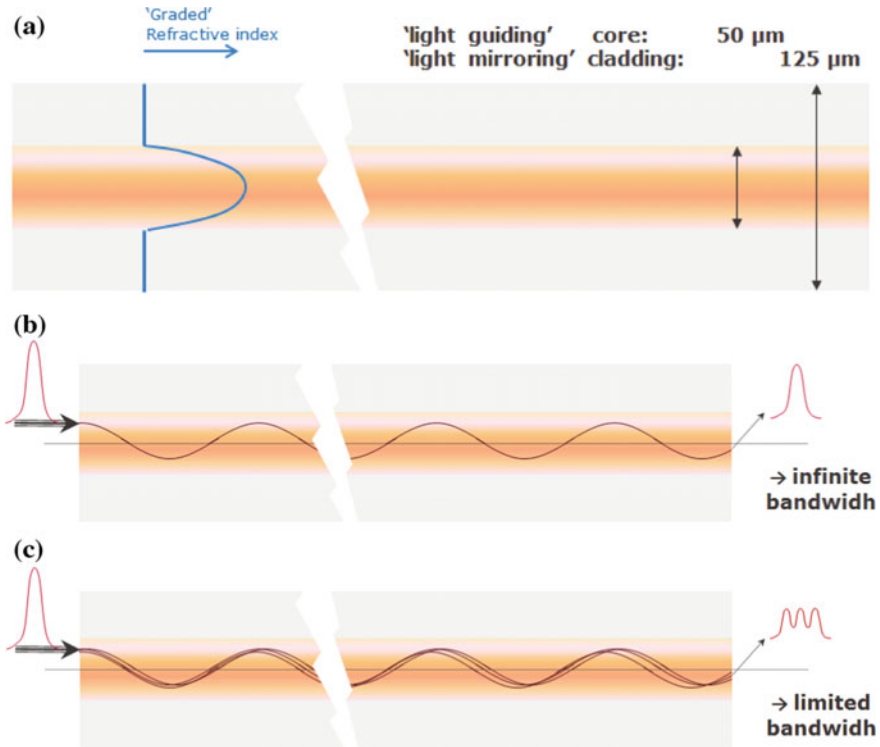


Fig. 3.2 Schematic representation of a laser pulse travelling through a graded index fiber (represented by **a**). In **b** the (as an example) three possible routes are equal, in **c** these differ slightly, resulting in pulse broadening at fiber output

are known, the calculated EMB is correct. Within the standardization documents, the characteristics of the laser are well thought-out.

The detected launch laser pulse should have a FWQM below a certain value. This value (ΔT_{pulse}) depends on three parameters: the spectral width of the laser ($\delta\lambda$), the chromatic dispersion of the fiber ($D(\lambda)$) and the length of the fiber (L). The lowest DMD value that can be measured with sufficient accuracy depends on ΔT_{pulse} . For OM4 fiber the lowest DMD specification is 0.10 ps/m. Figure 3.3 shows the relation between fiber length and minimum detector bandwidth as well as the minimum required value for ΔT_{pulse} (for the used laser and considered fiber: $\delta\lambda = 0.1$ nm, $D(\lambda) = 95$ ps/nm km). OM4 fibers are commonly used at a maximum system length (transmission distance) of 500 m. The length of the fibers that will be measured is chosen to be 550 m. The reason for this is practical; 50 m 'spare fiber length' to make a 500 m cable. DMD is measured on 'bare' fiber, so before the fiber is mounted into a cable structure. From Fig. 3.3 it shows the FWQM of the detected laser pulse must be below 60 ps, and the minimum detector bandwidth needed is about 15 GHz.

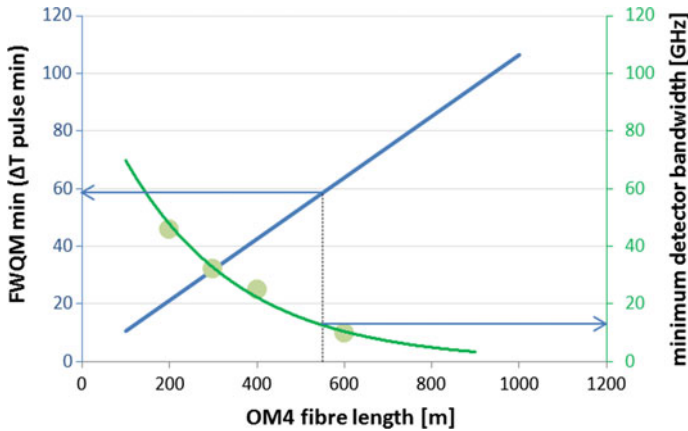


Fig. 3.3 Laser and detector requirements to allow qualification of OM4 fiber at different fiber lengths. For instance at fiber length 550 m, the minimum detector bandwidth should be about 17 GHz [9]

To accurately measure DMD below 10 ps/m on 550 m OM4 fibers, it makes sense to choose a ‘faster’ laser-detector combination, so a laser generating narrow pulses in time domain, and a very fast detector that can ‘follow’ these pulses in time. The fastest multimode detector on the market (2017) claims a bandwidth of 45 GHz. For the laser, things are not so easy, as the FWQM of the laser pulse is inversely related to the spectral width of the laser pulse. This time-frequency relationship limits the FWQM of the laser pulse to a minimum value. Briefly said, the spectral width of the laser may not be too large because then the effect of chromatic dispersion will reduce the resolution of the observed modal dispersion. Both dispersion effects occur, but the modal dispersion is what we are interested in. Chromatic dispersion is a material property, and will not differ that much from fiber to fiber, contrary to modal dispersion. The type of laser most commonly used to measure DMD on OM4 fiber is a ‘Picosecond’ Titanium Sapphire laser. The more commonly used ‘Femtosecond’ Titanium Sapphire laser generates very narrow pulses, but as a consequence the spectral width increases to values of a few nanometers. A maximum spectral width of 0.1 nm is preferred. For a Gaussian pulse shape this means a FWHM pulse duration near 10 ps is mandatory to measure DMD of OM4 fibers with a length of 550 m. The choice made for the laser configuration used for the experiments to follow: a 10 ps Picosecond Titanium Sapphire laser.

The requirements for the detector are not so well defined within the standardization documents. The speed of the detector is clear; the rise- and fall time must be sufficient to follow the laser pulse. Translated to detector bandwidth, for 550 m of fiber, a typical detector bandwidth of 20 GHz is needed. The other detector requirement is the ability to detect all modes leaving the fiber. This is not self-evident for fast detectors, as described in the introduction. This property is not specified by the detector vendor.

To characterize such a detector, a special designed optical fiber is used. Referred to as ‘mode separating fiber’. Regular multimode fibers are designed with an ‘Alpha Profile’ refractive index core [10]. If the Alpha value (α) is optimized for a particular wavelength, and the refractive index profile is very accurate, the EMB reaches very high values at that particular wavelength. This means all launched modes reach the detector at the same time after travelling through the fiber. The pulses are shaped exactly equal as launched into the fiber (assuming chromatic dispersion effects can be neglected). The refractive index of the fiber core usually is called ‘Alpha profile’ (α -profile), and is expressed as:

$$n(r) = n_1 \sqrt{1 - 2\Delta \left(\frac{r}{a}\right)^\alpha}$$

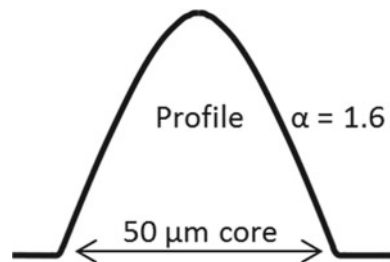
$$\Delta = \frac{n_1^2 - n_2^2}{2n_1^2}$$

where r is the distance to the center of the core, n_1 is the refractive index of the core at the center, n_2 is the refractive index of the cladding (surrounding the core) and a is the core diameter (50 μm). Solving the Maxwell’s equations leads to the dependency of bandwidth on the chosen value of α considered at the wavelength of operation (for OM4 this is 850 nm). A typical value for α is 2.06 [10]. If the fiber is produced with an exact correct α value over full fiber length, light propagates as schematically shown in Fig. 3.2b; the EMB of the fiber is infinite, all mode groups leave the fiber at exactly the same moment in time, and the pulse in time domain is relatively shaped equivalent as the pulse entering the fiber (only attenuation happens).

The specially designed fiber for purpose of detector characterization has an α of 1.6, and is referred to as ‘mode separating fiber’. The profile is shown in Fig. 3.4.

Because of the low α , but still used at 850 nm, the mode groups experience significant different ‘times of flight’ through the fiber. This can be seen clear on the resulting measured (normalized) DMD profile of Fig. 3.5b (laser is a Titanium Sapphire Mode Locked laser, pulses FWHM 10 ps, wavelength 850 nm). Figure 3.5a shows a DMD profile simulated on a fiber with an ‘ideal’ index profile, with an α value of exactly 1.6. The simulation model is described in [11]. The method to measure a fiber DMD profile is described in detail in *TIA* and *IEC* documents [6, 7]. The number of possible mode groups (18) reaches the end of the fiber at clear

Fig. 3.4 Core refractive index profile of the mode separating fiber; the α value is lower compared to conventional 50 μm graded index multimode fiber, resulting in a ‘narrower’ index profile [9]



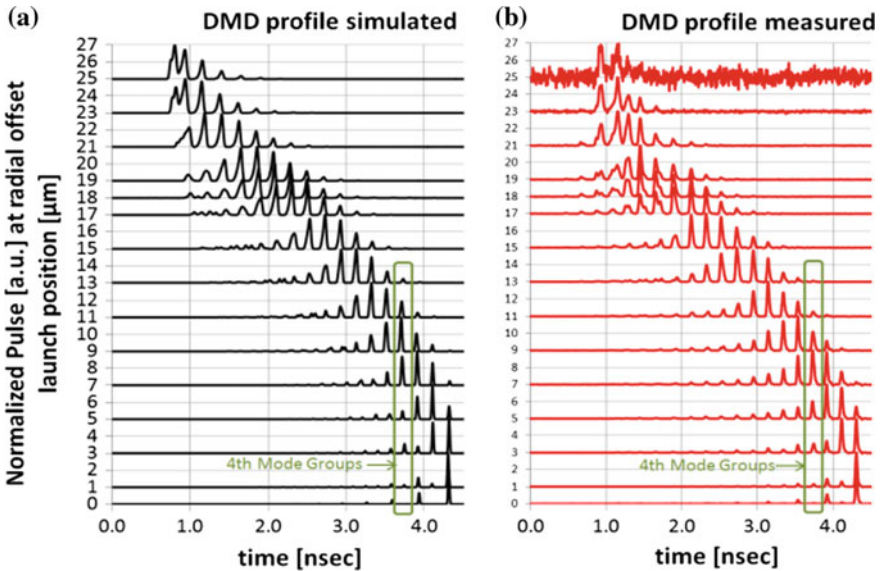


Fig. 3.5 DMD profile after 550 m of the mode separating fiber. The 18 Mode Groups are clearly separated in time. The 4th mode group is highlighted [9]

distinguishable moments in time. So one can derive the positions in time of the mode groups leaving the fiber, and also the power present within each of these mode groups. This can be done for all different radial offset positions. If the mode separating fiber is ‘perfect’ (ideal α -profile, and very homogeneous over fiber length), the difference between the measured and simulated DMD profile is a good indicator for detector performance. The more the DMD profiles look similar to each other, the better the performance of the detector (it then truly detects all mode groups leaving the fiber). The length of the mode separating fiber is 550 m.

3.3 Mode Separating Fiber

In Graded Index (GI) multimode fibers modes are sorted into mode groups. Modes within the same group ‘see’ nearly the same group index. And also the same group index difference with its neighboring group. So time of flight differences are observed only for modes that are not in the same group. The difference of time of flight between consecutive mode groups is at first order determined by the alpha value (α), the numerical aperture (NA) or Delta (index difference core-cladding) [5], the core diameter (a) and the operational wavelength (λ). And at second order a function of the dopant content (full Germanium and fluorine or a Germanium and Fluorine co-doping).

For the mode separating fiber, the α value is chosen so that the mode groups experience a distinguishable different time of flight when travelling through the fiber (chosen length: 550 m). This condition can be expressed as follows:

$$\frac{|\Delta\tau| \cdot L}{\Delta T_{\text{REF}}} > X$$

where $\Delta\tau$ is the time of flight difference between consecutive mode groups in ps/m, L is the minimum useable fiber length in m, ΔT_{REF} is the Full Width Quarter Max (FWQM) of the ‘reference’ pulse used in the DMD measurement in ps (note ΔT_{REF} equals ΔT_{pulse} if chromatic dispersion through the fiber is negligible, which is the case when the pulse travels through short fiber lengths).

X is a ‘threshold’ value that is larger than 4. For a 50 μm diameter core GI multimode fiber with an NA = 0.200, one can approximate $\Delta\tau$ as follows:

$$\Delta\tau(\lambda, \alpha) = P_{00} + P_{10} \cdot \lambda + P_{01} \cdot \alpha + P_{11} \cdot \lambda \cdot \alpha + P_{02} \cdot \alpha^2$$

The constants P_{xy} are computed using the simulation model described in [11]. The result is plotted in Fig. 3.6a. Figure 3.6b shows the pulses of two consecutive mode groups leaving the fiber, referring also to Fig. 3.5. When $\alpha = 1.6 \rightarrow \Delta\tau = 0.36$ ps/m at 850 nm, so the delay between 2 mode groups after 550 m fiber is 200 ps.

The theoretical power distribution over core radius of pulses leaving the fiber at DMD launching conditions is shown in Fig. 3.7. While approaching the edge of the core (25 μm), the power will go down; some pulse energy is launched in the cladding. In practise power is also lost because of core-cladding interface artefacts and slight fiber bending. The power distribution over the core radius (called ‘Shape’) depends

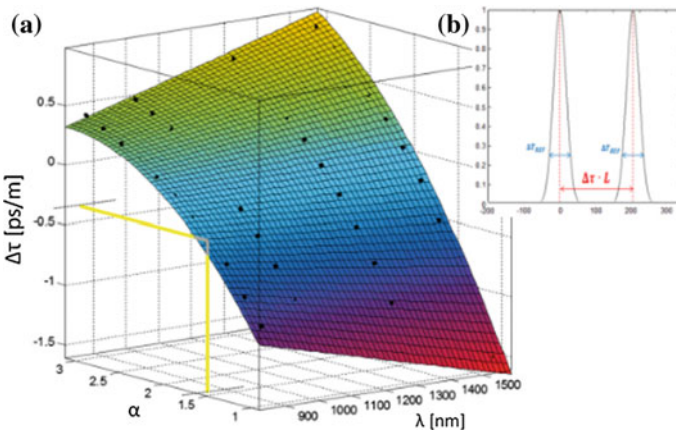


Fig. 3.6 **a** 3-D plot showing the theoretical relation between the time delays between consecutive mode group ($\Delta\tau$), the wavelength (λ) and the α value of the fiber core index profile. **b** Time of flight difference two consecutive mode groups [9]

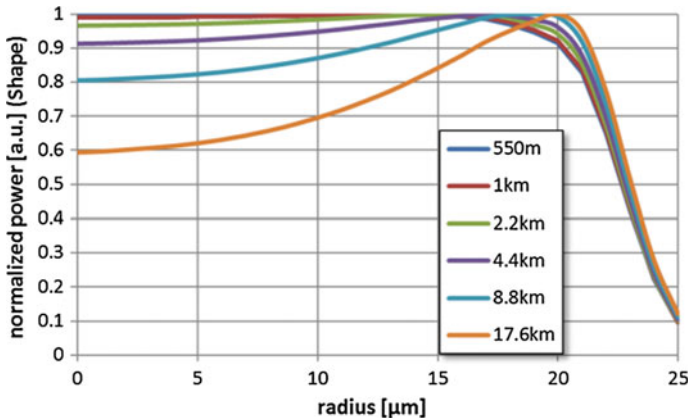


Fig. 3.7 Fiber length dependent (theoretical) power distribution ('Shape') over a Graded Index 50 μm core fiber. For short fibers, the power is constant up to typical 17 μm offset launch radius [9]

also on the length of the fiber. For short fibers the power is constant up to typical 17 μm , for longer fibers a 'dip' pops up. This is caused by the higher Germanium content at the centre of the core.

The mode separating fiber is made by using the Plasma Chemical Vapor Process (PCVD) [10]. A 'sliding' cylindrical resonator creates a plasma inside a glass tube, causing thousands of very thin layers of glass deposited to the inner side of the tube. This continues until a very small channel is left. This channel is closed in a later stage by collapsing the rod. The refractive index of the deposited glass changes with content of the gasses flowing through the tube. Each layer receives slightly more Germanium, causing the refractive index to increase according to the targeted 'Alpha-profile'. The resulting rod is drawn to fiber, the glass diameter close to 125 μm , added are 2 coating layers to protect the fiber. The fiber is checked by a Exfo NR-9200 high resolution Optical Fiber Analyzer. A small laser probe scans the flat cleaved fiber interface in two directions, the refracted beam is detected and refractive index is computed. The results are shown in Fig. 3.8:

The value of α and a are fitted using the formula for $n(r)$. The fitted value for α is 1.6200, the value for a is 26.5096 μm . The variation of the index profile from the fitted trace (Fig. 3.8b) is quite small, so a good match between theoretical and experimental DMD data is expected, and shown in Fig. 3.5.

To give an idea; consider the pulse launched at 9 μm from the center of the core. Using the model to predict the shape of the pulse train leaving the fiber results in Fig. 3.9. An α difference of 0.02 results in a difference of time of flight for the (for instance) 7th mode group of 0.15 ns/km. An α difference of 0.00002, typical seen from Fig. 3.8b, means a difference of 0.15 ps/km, so 0.08 ps over 550 m fiber. This is below the DMD measurement resolution.

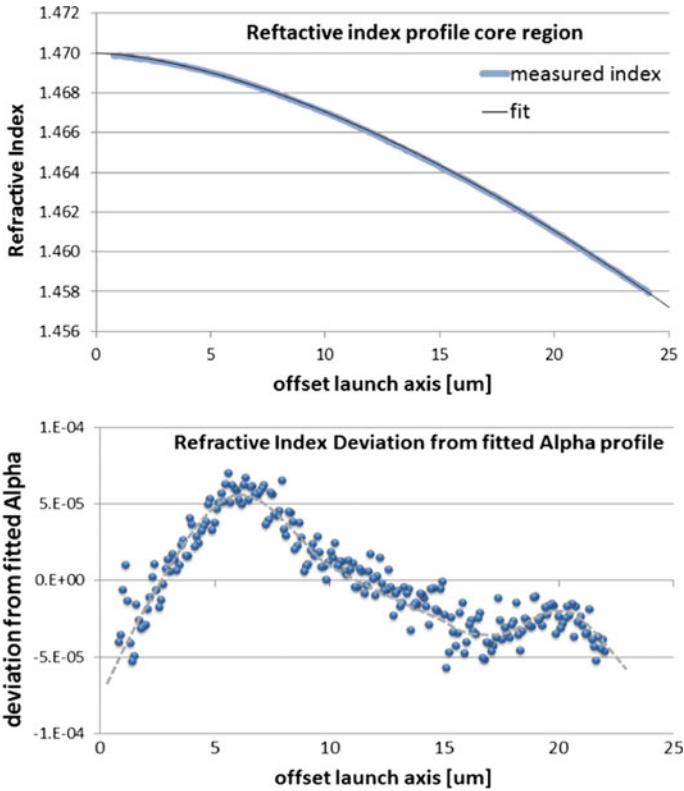


Fig. 3.8 Measured and fitted core refractive index profile of the mode separating fiber

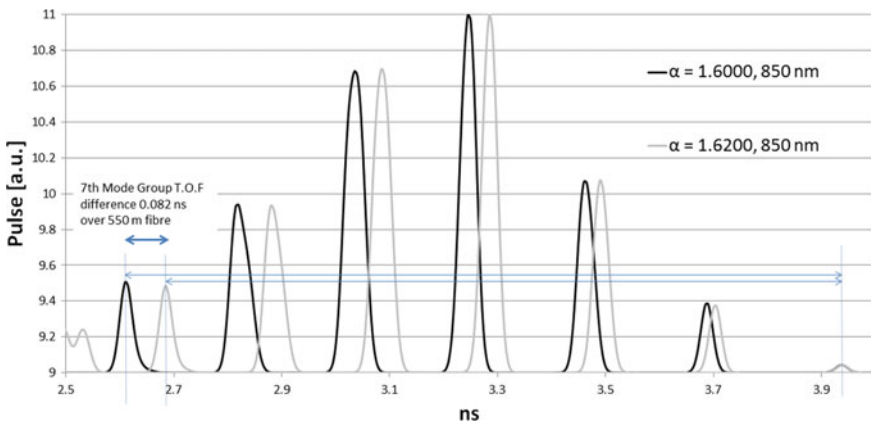


Fig. 3.9 Launched at 9 μm radial offset, theoretical pulse train leaving 550 m of fiber supposing an α difference of 0.02 exists (1.62 vs. 1.60). The time of flight of the 7th mode group is 0.15 ps/m (most right pulse is the 1st mode group)

To summarize: the strong point of the mode separating fiber is an accurate index profile with very low variation over the fitted α value. The weak point is an unknown variation over fiber length of the fitted α .

3.4 The Detectors

In total five commercially available detectors are tested. The goal is not to test these detectors, and promote the best performing detector. The goal is to compare these high speed multimode detectors, and discuss to what degree these detectors can be used to qualify OM4 fiber. The targeted fiber measurement length is 550 m (the maximum length in systems is 500 m). The five detectors are all classified as high speed optical electrical converters, fiber coupled, FC-PC connected. The detectors are available on the market before and until the year 2017. The detectors will remain anonymous. The exact internal structure of these detectors is unknown. The table shows the five detectors and their main specified or (occasional) measured characteristics. The indication ‘tapered’ may also be ‘lensed’ (the way the internal pigtail is treated to push the light on the smaller detector sensitive area) (Table 3.1).

Detector #1 is an ‘integrated optical-electrical module’ (responsivity in A/W is not specified). Detectors #2 and #5 use an internal tapered (or lensed) pigtail; the sensitive detection area is estimated to be near 30 μm diameter. Detector #3 is kind of special, because it is specified very fast (45 GHz), despite the fact that a ‘direct’ coupling exists of the 50 μm internal pigtail to the sensitive detection area (so the internal fiber pigtail is *not* tapered or lensed). The relatively large area must be thin to reach the high detection speed, and as a consequence the responsivity is very low compared to the other detectors [8].

From Fig. 3.3, detector #1 is just too slow (10 GHz), case an OM4 fiber is measured at length 550 m. The specified detector bandwidth should then at least be 14 GHz. But still, detector #1 is taken along.

Table 3.1 The five commercially available high speed detectors with specifications. The indication ‘tapered’ may also be ‘lensed’ (this is not explicitly specified by the manufacturer) [9]

nr	Specified range (nm)	Responsivity A/W @ 850 (nm)	Specified optical bandwidth (GHz)	Internal fibre pigtail
1	700–1650	?	10	62.5 μm ‘?’
2	700–1650	0.59	20	62.5 μm ‘tapered’
3	400–870	0.08	45	50 μm ‘direct’
4	700–890	0.4	30	50 μm ‘?’
5	800–1650	0.48	25	62.5 μm ‘tapered’

3.5 Measurement Results

3.5.1 Reference

First objective is to verify the theoretical power distribution as shown in Fig. 3.7. The most straightforward way to verify this is using an optical detector with a homogeneous sensitive area significantly larger than the core of the fiber ($50\ \mu\text{m}$). So all modes leaving the fiber will be detected. The guided light output of the fiber must be radiated straight on the sensitive area of the detector, no optics (or fiber) in between the fiber output and the sensitive detector area. Detectors with a large sensitive area are much too slow to follow the rise and fall of the launched laser pulses, so a DMD profile like shown in Fig. 3.5 can't be measured. However the power leaving the fiber at each launch position can be detected by modulating the launched beam at a low frequency (chopper), and use a Lock In amplifier to detect the power.

The measurement is done by performing a 'DMD scan' in 4 directions (4-Quadrant scan: 4Q), and measure the power at each launch position. The way to do a 4Q DMD scans is well described in the standardization documents [6, 7]. Summarized in Fig. 3.10.

The laser source is the aforementioned 10 ps Picosecond Titanium Sapphire laser (Ti:Sapp). The 4Q scan enables accurate alignment of the fiber, so the launch at $0\ \mu\text{m}$ radial position indeed is at the optical center of the fiber. After the DMD scan, the 4 quadrants are combined, for instance by folding the pulses at each launch radius. The pulses leaving the fiber normally go to a high speed detector connected to (for instance) a sampling module and signal analyzer. The large area detector is too slow to follow the fast Ti:Sapp laser pulses, so the incoming beam is modulated with a chopper at 160 Hz, connected to a Lock In Amplifier. At each radial offset launch position the signal is measured. Finally the 4 quadrants are averaged, resulting in the graph shown in Fig. 3.11: showing the experimental power distribution and the theoretical power distribution that was already shown in Fig. 3.7.

At the more distant launch positions ($23\ \&\ 25\ \mu\text{m}$ from the center), the experiment shows higher power relative to theory. We believe this difference is caused by artifacts

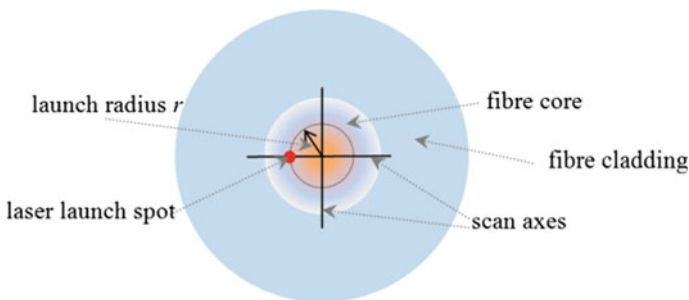


Fig. 3.10 4Q DMD scan, laser launch spot size is $5\ \mu\text{m}$ diameter

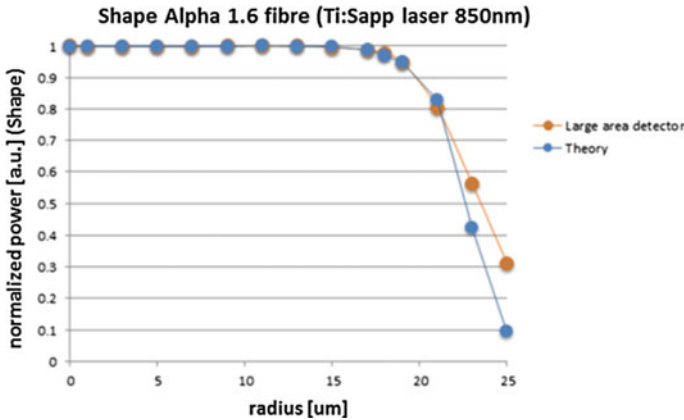


Fig. 3.11 Theoretical and experimentally verified power distribution ('Shape') of the mode separating fiber. The power is constant up to typical 17 μm offset launch radius [9]

in the fiber index profile near the core-cladding interface (scattering, leaky modes). Some power might travel through a shallow channel in the cladding region close to the core. Up to 21 μm theory and experiment are in close agreement. Relevant is the radius where the power starts to drop ($\sim 17 \mu\text{m}$), this happens for both theory and experiment, and is a good indication of the accuracy of the launch positions.

3.5.2 Detectors Under Test

The coupling of laser pulses into the mode separating fiber is realized by a direct coupling (no in between optics) of a HP780 launch fiber (single mode at 850 nm, launch spot diameter at 850 nm is 5 μm) to the mode separating fiber. For both fibers, the cleave quality is checked by an interferometric technique, and is far below an angle of 1° (to avoid angular coupling to the mode separating fiber). The 4Q DMD scan (including alignment, 850 nm) is executed five times per detector to visualize the repeatability of the measurement. For detectors #3 and #4 the five power distributions are shown in Fig. 3.7a. The responsivity of detector #3 is very low, causing this detector to have the poorest reproducibility. Figure 3.12b shows the averaged power traces of all five detectors.

It is evident none of the detectors approach the theoretical Shape. None of these detectors capture all the modes leaving the fiber. Since the internal structure of the detectors is unknown, it is not possible to explain the observed differences.

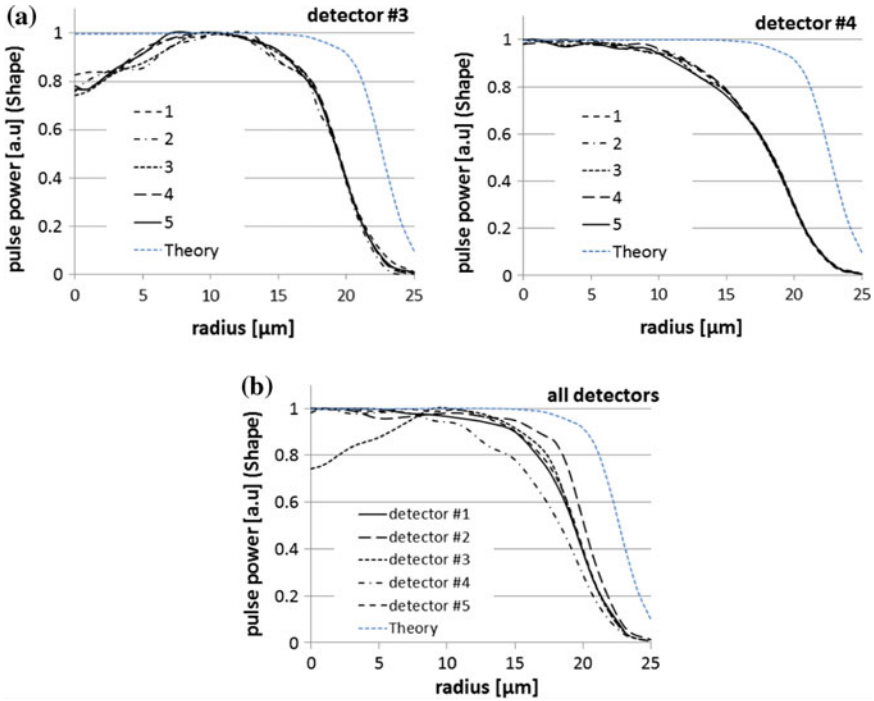


Fig. 3.12 **a** 5 pulse power distribution measurements on detectors #3 and #4. **b** Averaged 5 pulse power distributions of the five detectors [9]

3.5.3 Discussion

The fiber pigtailed inside detectors #2 and #5 are ‘tapered’ or ‘lensed’. So the fiber output field to the detector sensitive area is ‘pushed’ to a smaller area (from 62.5 to approximately 30 μm diameter). Not all modes reach the sensitive area (or do reach the area, but fail to generate current). It is expected the lost modes are the modes that travel at the edge of the core. The power distribution (Shape) for both detectors is nearly equal, detector 2 reaches the theoretical Shape a little closer (detector 2 has a slightly lower bandwidth, so maybe the sensitive area is just a little larger).

Detector #1 shows an equivalent Shape to #2 and #5, which could mean it is also constructed with an internal tapered or lensed fiber pigtail.

Detector #3 is the ‘oddball’ of the selection. It has an internal fiber with a 50 μm diameter core. This fiber pigtail is *not* tapered or lensed, and -as a consequence-, uses a larger and likely also thinner sensitive area. So it is specified with a significantly lower responsivity. Still, the Shape does not approach the theoretical curve. Two possible explanations for the narrower Shape are possible:

- (1) Bend losses of the internal pigtail (50 μm shows more bend loss compared to 62.5 μm), as the pigtail is bended over few cm inside the detector housing.

- (2) A poor coupling of the pigtail to the sensitive area (pigtail is glued to the sensitive area with a slight mismatch).

To overcome (1), an interesting solution might be to use ‘bend-insensitive’ 50 μm fiber as the internal pigtail.

The ‘dip’ in the middle of the Shape of detector #4 is harder to explain. Possibly the homogeneity of the sensitive detector area is poor because it needs to be thin to reach the high speed (45 GHz). The only way to verify this is to inspect the sensitive detector layer, which is a detector destructive test.

Finally detector #4. The Shape is kind of ‘roundish’. The size of this detector is very small, and the housing is completely sealed. The internal structure is unknown (not even sure an internal pigtail is inside). This detector shows the most narrow shape of all five detectors.

So which detector is most suited to qualify OM4 fiber? The bandwidth of detector #1 is just below the minimal required value to measure 550 m fibers. Detector #3 fails to detect all power of the inner mode groups, this is absolutely unacceptable. As in systems most power is launched to the center of the core, and so in this region the pureness of the fiber index profile is very relevant. So all modes travelling through the center must be detected. Detector #4 shows the narrowest (roundish) shape, so deviates more from theory than detectors #2 and #5. Detector #2 is closest to theory, but Detector #5 shows a higher bandwidth. Either Detector #2 or Detector #5 is the best choice.

3.5.4 A Closer Look

The results based on the power distribution will show as well by using a regular OM4 fiber (α close to 2.06) rather than the mode separating fiber (α close to 1.62). The theoretical differences in Shape are almost independent from the α value. Using the mode separating fiber however creates the possibility to verify the detector response not only per launch offset radius, but also per mode group. To simplify the analysis, we only consider detectors #2 (widest Shape), #3 (center dip) and #4 (narrowest Shape). Figure 3.13 shows the power levels at arbitrary units of the 18 mode groups leaving the mode separating fiber. The detected power levels per mode group at 4 different launch offset radii are plotted alongside the theoretical values computed from the simulated DMD profiles (as shown in Fig. 3.5a).

To scale the power levels equal for comparison (theory vs. experiment), it is assumed at 9 μm radial launch offset, all fiber guided mode groups are detected (based on the results shown in Fig. 3.12b). Checking Fig. 3.5, it shows the equivalence between the simulated and measured DMD profile is fair. However when approaching the cladding, the differences seem to increase (the measured data belongs to detector #2). This could be caused by core-cladding interface artifacts of the mode separating fiber. Further, to optimize the model, one must know the exact Alpha value (α), core-radius (a) and Delta (Δ) [10, 11]. At one position (the ‘top-side’ of the fiber)

these were measured by a Exfo NR-9200 high resolution Optical Fiber Analyzer, but at other positions these parameters may be different. The process of fiber making is optimized for ‘regular’ OM4 fiber with an α value near 2.06, for intentional ‘specialty’ fibers like the mode separating fiber, the homogeneity over fiber length will be less good.

From Fig. 3.13a, theory, ‘odd’ mode groups are symmetric whereas ‘even’ mode groups are anti-symmetric [11]. So no power at fiber output by the even mode groups (2, 4, ...). This is well confirmed by the experimental data.

Detector #3 fails to detect full power of the first mode group, while the third and fifth mode group approach theory. Likely caused by an artifact in the center of the detector sensitive area, which is a reason to reject detector #3 for OM4 fiber qualification.

Figure 3.13b, c (launch radial offset 9 and 18 μm) show a typical but unexpected result. The measured mode group power distributions of the three detectors drift to higher order modes compared to theory. One might expect the opposite, as higher order mode groups are more sensitive to selective mode detection because these travel mainly closer to the cladding.

This means even so the total detected power per radial offset launch position is in good agreement with theory (Fig. 3.11), the distribution of this power over the mode groups is not. So how can we explain this observation? Macrobend induced losses (because the fiber is on a smaller spool) are excluded, because at 850 nm, this will impact only the most outer mode groups (16–18). Another fiber aspect could be tension on the fiber. Also the impact of profile inhomogeneity’s on the detected power distribution over the mode groups needs further studies.

Figure 3.13d, when launching close to the cladding, it clearly shows the loss of power for all three detectors compared to theory. This is the main cause of the narrow experimental Shapes shown in Fig. 3.12b.

3.6 Conclusions

As far as we know today, no high speed optical detector commercially exists that is capable of detecting all modes leaving a 50 μm graded index fiber. For 4 detectors the indirect cause of this observation is the fact that the sensitive detector area is smaller compared to the fiber core area. Inside the detector the fiber guided light is ‘pushed’ to a smaller area, causing power not to reach the sensitive area (or does reach it, but does not generate current). The mechanism that causes this effect can only be explained once the exact internal structure of the detector is known. For instance the characteristics of the internal fiber pigtail. From experiments it is clear most of the ‘lost’ power belongs to mode groups travelling at the edge of the fiber core. However a drift of power to higher order mode groups is observed for offset launch radii already few micrometers from the center of the core. This observation needs further studies, is it a slight mismatch between predicted mode group propagation through the fiber and the experimental data, or is this effect caused inside the detector.

For purpose of OM4 fiber qualification, detectors #2 and #5 will be the best choice. These will not detect all modes leaving the fiber. Mainly the modes travelling close to the edge of the core do not contribute to the EMB computation. This does not mean that the measured EMB is by definition too low, and the ultimate influence on the system speed is expected to be very limited, as not much power travels at the edge of the core.

A first optimization to the internal structure of the detector might be to use a 'bend-insensitive' GI multimode fiber to serve as internal pigtail, to replace the 'conventional', non bend-insensitive OM1 (62.5 μm) or OM2 (50 μm) fiber. This will cause fewer modes to leave the bended internal fiber pigtail from the side.

References

1. CommScope, Wideband Multimode Fiber - What is it and why does it makes sense? (White paper) (2015). http://www.commscope.com/docs/wideband_multimode_fiber_what_why_wp-109042.pdf
2. D. Molin, F. Achten, WideBand OM4 multi-mode fiber for next-generation 400Gbps data communications, in *IWCS* (2014)
3. R. Pimpinella, B. Kose, J. Castro, Wavelength dependence of effective modal bandwidth in OM3 and OM4 fiber and optimizing multimode fiber for multi-wavelength transmission, in *IWCS* (2014)
4. M. Bigot, D. Molin, Wide-band OM4 multimode fibers for future 400Gbps and 1.6Tbps WDM systems, in *IWCS* (2015)
5. K. Oh, U. Paek, *Silica Optical Fiber Technology for Devices and Components (Chapter Five)*, (Wiley, 2012)
6. TIA, FOTP-220 - differential mode delay measurement of multimode fiber in the time domain - TIA-455-22-A (2003)
7. IEC, IEC 60793-1-49 ED3 - optical fibers - Part 1-49 (Draft): measurement methods and test procedures - differential mode delay (2017)
8. R. Hui, M. O'Sullivan, *Fiber Optic Measurement Techniques* (Chap. 1, p. 35) (Elsevier Academic Press, 2009) (2012)
9. F.J. Achten, D. Molin, Characterization of high speed optical detectors for purpose of OM4 fibre qualification: selective mode detection, in *Proceedings of the 6th International Conference on Photonics, Optics and Laser Technology (Photoptics 2018)* (2018), pp. 141-147
10. H. Venghaus, N. Grote, *Fibre Optic Communication* (Chap. 2). Springer Series, 2012 (2017)
11. A. Gholami, D. Molin, P. Sillard, Physical modeling of 10 GbE optical communication systems. *J. Light. Technol.* **29**(1) (2011)
12. D. Molin, M. Bigot, F. Achten, A. Amezcua-Correa, P. Sillard, 850-950 nm WideBand OM4 multimode fiber for next-generation WDM systems, in *OFC* (2015)

Chapter 4

Polarization Modulated Vertical-Cavity Surface-Emitting Lasers in Quantum Key Distribution



Ágoston Schranz and Eszter Udvary

Abstract Vertical-cavity surface-emitting lasers (VCSELs) have multiple beneficial properties in quantum key distribution (QKD). However, polarization switching (PS), which happens between two orthogonally polarized eigenmodes, is characteristic of these lasers. PS is extremely problematic if it is unwanted in all polarization-sensitive applications. The origin and properties of PS is discussed, along with potential error scenarios introduced to QKD protocols BB84 and B92. We propose a new transmitter design for the BB84 protocol using only two VCSELs—both corresponding to one of the two bases in which polarized photons are sent—, which are modulated in polarization, purposely generating switches between two orthogonally polarized modes. The methods of polarization modulation are described, along with advantages and design difficulties of the new design. We also consider the possibility of a spectral attack performed by an eavesdropper, originating from the frequency splitting between the polarization eigenmodes of a single VCSEL, and offer a theoretical solution that can protect the key from eavesdroppers utilizing this kind of attack.

4.1 Quantum Key Distribution

In this chapter, we are going to introduce and propose a novel transmitter design for one of the simplest quantum key distribution protocols. As a starting point, Sect. 4.1 gives a brief introduction to the motivations behind our work by shortly describing the objectives of cryptography and QKD, as well as the practical implementation difficulties concerning these fields.

Á. Schranz (✉) · E. Udvary
Budapest University of Technology and Economics, Budapest, Hungary
e-mail: schranz@hvt.bme.hu

E. Udvary
e-mail: udvary@hvt.bme.hu

© Springer Nature Switzerland AG 2019
P. Ribeiro and M. Raposo (eds.), *Optics, Photonics and Laser Technology 2018*, Springer Series in Optical Sciences 223,
https://doi.org/10.1007/978-3-030-30113-2_4

4.1.1 *Necessity of QKD in Cryptography*

The need for cryptography has been present throughout history. Nowadays, most cryptographic schemes use asymmetric public-key algorithms, which do not require the two parties to agree on a pre-shared secret. Every party has two keys, one of them made public, the other being private. Messages encrypted by one's public key can only be decrypted with the help of the private key, therefore unauthorized people cannot access its content.

The security of all these public-key algorithms depends on the computational complexity of mathematical problems, for example prime factorization of large numbers in case of the RSA scheme. As soon as general purpose multi-qubit quantum computers will be available, the time taken to solve these problems is going to be reduced drastically. This means that the security of public-key cryptography will be breached, and other, provenly secure algorithms should be implemented in practice.

In a one-time pad scheme, keys have to be at least as long as the message waiting to be encoded. The message and key characters are then paired and combined by modular addition to generate the ciphertext character. If messages and keys are binary, modular addition simply reduces to taking the bitwise exclusive OR (XOR) of pairs. However, every single key should only be used once to avoid information leakage, therefore a new key should be agreed on before every transmission. The security of one-time pad schemes (namely, that the cyphertext does not provide any information about the original message, apart from its maximum length, in case of a truly random key), was proven by Claude Shannon [1].

The main problem concerning the one-time pad (or any other symmetric-key algorithm) is that the keys have to be secretly shared between the communicating parties before the transmission of encrypted messages can begin. Using a classical channel for key distribution provides eavesdroppers an easy opportunity to gain access to a third copy of the key, while physical transportation and storage of keys in "sealed envelopes" or on any other data storage devices is impractical in the long run.

Quantum key distribution—maybe the most advanced field of quantum communications—offers provably secure solutions to key distribution. The security is based on the underlying principles of quantum mechanics, causing any disturbance induced by the eavesdropper (referred to as Eve in many texts) to alert the sender and the receiver (often referred to as Alice and Bob, respectively). Alice and Bob will then abort the key transmission and restart it, therefore Eve will be unable to gain information. As a result of this, QKD protocols do not offer protection against communication suppression, so an always active eavesdropper can successfully block the transmission of encrypted messages.

4.1.2 QKD Protocols and Practical Implementations

The distinctive features of quantum mechanics utilized by QKD protocols are mainly the Heisenberg uncertainty principle and the no-cloning theorem [2]. The latter states, that there exists no single operation on quantum objects which is able to perfectly copy any given quantum state, only if the input state is known a-priori or all the possible states are orthogonal to each other. As a consequence, the quantum states representing qubits used in QKD are not all pairwise orthogonal to each other (though some of them may be).

The linear polarization of a single photon is often used as a qubit in discrete-variable quantum key distribution (DV-QKD) protocols, for example BB84 [3], B92 [4] or E91 [5]. Although this qubit choice is robust for communication purposes, the implementation of true, on-demand single photon sources is a difficult challenge. As a substitute, the practical devices use weak coherent states, approximated by highly attenuated laser pulses [6]. All semiconductor laser types, including vertical-cavity surface-emitting lasers as well as conventional edge-emitting lasers (EELs) are suitable for weak coherent state QKD schemes.

Coherent states follow significantly different photon statistics compared to number or Fock states: while a Fock state has a well-defined number of photons (one photon in case of a single-photon state), coherent states have Poissonian photon statistics [7]. This means that the probability of measuring n photons in the state is given by (4.1).

$$P(n) = e^{-\lambda} \frac{\lambda^n}{n!} \quad (4.1)$$

Coherent states are thus characterized by the expectation value (mean photon number) λ of the distribution, which is analogous to the classical light intensity. As a Poisson distribution with any positive parameter λ has nonzero values for any integer n , a weak coherent state will sometimes contain one photon, sometimes more photons, and in some cases none at all. Decreasing the mean photon number will lead to more blank states, effectively reducing the achievable key rate, since these states carry no information. On the other hand, increasing the expectation value will increase the probability of multiphoton states, which allow the eavesdropper to perform a photon number splitting attack [8] that greatly reduces the security of the protocols. This trade-off is often solved by using states with mean photon numbers around $\lambda \approx 0.1$. In this intensity regime, the probability to detect no photons in a light pulse is above 90%, leading to a key rate at most one tenth of the laser's pulse repetition rate, but only a very small percentage of pulses will contain multiple photons.

4.1.3 Examples of Simple DV-QKD Protocols

For a better insight into the principles of QKD, how some of the simplest protocols work and why they provide security against eavesdropping, the operation of two basic protocols, the original BB84 and B92 are described in the following.

BB84: The First QKD Protocol. In 1984, Charles Bennett and Gilles Brassard described the first quantum key distribution protocol in their article *Quantum cryptography: Public key distribution and coin tossing* [3], generally referred to as BB84 after the authors' initials and the year it was published. Although more sophisticated protocols have been invented since then, BB84 still serves as a useful starting point to study QKD due to its relative simplicity and clear principles of operation. Moreover, its implementation is also relatively simple, making it a good candidate for practical realization. The following explanation is given in terms of single photon polarization qubits.

BB84 uses four different quantum states. Two of these form the rectilinear basis, the other two form the diagonal basis. Both are bases of a two-dimensional space of possible polarizations, rotated by 45° relative to each other. Before every transmission, Alice generates two independent random bits, s_A and m_A . s_A is the key bit to be transmitted, m_A decides the basis in which the key bit is going to be encoded. In the rectilinear basis, $s_A = 0$ is encoded onto a horizontally (\rightarrow), $s_A = 1$ is encoded onto a vertically (\uparrow) polarized photon; in the diagonal basis, $s_A = 0$ results in a photon with polarization angle $+45^\circ$ (\nearrow), while $s_A = 1$ corresponds to -45° (\searrow).

On the receiver side, Bob initially has no information about either s_A or m_A , so he generates a random bit m_B to decide about the measurement basis. If he chooses correctly ($m_A = m_B$), for example, measuring a vertically polarized photon's polarization in the rectilinear basis, he will always get the correct measurement result s_B . Was he to choose the incorrect basis, however, the projective measurement would result in a completely random value for s_B . After transmitting all the key bits, the two parties disclose their basis choices m_A, m_B on a classical channel, keeping only the key bits where these choices agree, discarding everything else. This ensures that—using a perfect, error-free quantum channel—the remaining raw keys $\{s_A\}, \{s_B\}$ are equal in their respective positions.

If an eavesdropper (Eve) decides to try and gain some information about the secret key, she needs to behave like Bob and measure the polarization in a random basis. This disturbs the quantum state and leaves a trace: measurement in the wrong basis causes the wave function to collapse into the measured state, effectively altering the qubit. This would mean that even if Bob chose the correct basis, some of the measurement results would differ from what Alice intended to send. Fortunately, they do have a way to discover this hazard, called key sifting: by sacrificing and comparing some random set of the raw key bits, they can check whether the error rate exceeds the threshold calculated from the channel errors and the effects of eavesdropping. If the error rate is too high, they abort the transmission and restart it with a fresh key distribution, ensuring that Eve gains no information about their shared secret.

B92. Eight years after *BB84* had been published, Charles Bennett described in his paper a protocol that is now known as *B92* [4]. *B92* only uses two non-orthogonal states, \rightarrow and \nearrow . For every transmission, Alice generates a random bit a . Whenever $a = 0$, she sends a horizontally polarized photon, however, if $a = 1$, she sends a photon polarized along $+45^\circ$. Note that both of these states correspond to logical zeroes in *BB84*. On the receiver side, Bob generates a random bit a' before every measurement and measures the photon's polarization in the rectilinear or diagonal bases for $a' = 0$ and $a' = 1$, respectively. The measurement results b are interpreted the same way as in *BB84*: photons with a measured polarization angle 0° (\rightarrow) or $+45^\circ$ (\nearrow) result in a logical zero, those with polarization angles of 90° (\uparrow) or -45° (\searrow) result in logical ones.

If he chooses the correct basis ($a = a'$), the measurement result will always be zero. Measurements of ones are only possible if the basis choices differ ($a \neq a'$). In the latter case, the results are completely random and unbiased, so we can expect to get an outcome of 1 in 25% of all measurements. After all bits are transmitted, Bob discloses his b values on a classical channel. The communicating parties keep their respective a and a' values for all $b = 1$ and discard those where $b = 0$. The sequence of a s and a' s that are not discarded form the raw keys, but it has to be taken into account that they are the negated versions of each other ($a = \bar{a}'$). The key sifting part of the protocol is similar to that in *BB84*.

4.2 VCSELs in QKD Systems

As a central point of this discussion, in Sect. 4.3 we propose a new transmitter design for the *BB84* protocol that uses vertical-cavity surface-emitting lasers. To justify this proposition, let us first describe the main properties of VCSELs, focusing mainly on those that are relevant in the framework of quantum key distribution.

The following sections explain why VCSELs are valid contenders for QKD applications compared to edge-emitting lasers, but also give a detailed overview of a mechanism potentially dangerous to some of the key distribution protocols: polarization switching.

4.2.1 Advantages of VCSELs over EELs in QKD

VCSELs have inherent advantages over edge-emitting lasers in most low-power applications, such as DV-QKD protocols operating in a near single photon power regime. This originates from the fact that VCSELs typically have lower threshold currents and lower output power compared to their edge-emitting counterparts [9]. However, to reach the single photon level, strong external attenuation is necessary regardless of the light source in use. In case of VCSELs, this need results in less attenuation and power waste than for EELs, offering a more environmentally friendly, energy efficient solution.

Due to their nature, single photon DV-QKD implementations are very sensitive to power fluctuations. Additional losses should be kept as low as possible, so that the output mean photon number can be well-controlled by the external optical attenuators. A good example is the case of coupling losses from the light source to the optical fiber: manually disconnecting and reconnecting the fiber can produce different outcomes in terms of coupling efficiency. VCSELs' output light beams have a circular cross section owing to their highly symmetric structure, also characterized by low divergence angles. These properties lead to more efficient, lower loss fiber coupling, but can also be utilized in long free-space QKD links, because EELs typically have elliptical beams with higher divergence angles [9]. Given these advantages, there already exist practical QKD devices which use VCSELs as photon sources [10].

4.2.2 Drawback: Polarization Switching

Some of the VCSELs exhibit a very specific mechanism that is not found in properly constructed edge emitting lasers: polarization switching. Unlike in an EEL, the surface-emitting laser's design is highly symmetric, which in itself leads to no preferred linear polarization direction. The circular symmetry is, however, never perfect owing to e.g. manufacturing processes and crystallographic structures, manifesting in small, inevitable anisotropies within the device. These imperfections mostly choose two preferred polarizations for the output light (polarization eigenmodes) that are orthogonal to each other and correspond to the crystallographic axes. Above threshold current, one of these modes is selected and starts to lase, suppressing the other. By increasing the injection current, an abrupt switching to the orthogonal mode—still while staying in the fundamental transverse mode current regime—might be observed (Fig. 4.1). Due to the phase anisotropy (birefringence), the two eigenmodes have a frequency split between them. Switching from the high to the low frequency mode with increasing current is referred to as Type I PS, the reverse is called Type II PS. A given VCSEL might show only Type I PS, Type II PS, both of them or none at all: it can vary even within a batch of lasers coming from the same wafer. Apart from the injection current and anisotropic behaviour, polarization switching strongly depends on the strength and polarization angle of external optical feedback into the VCSEL cavity [11].

The underlying reasons responsible for polarization switching in VCSELs have been extensively studied over the years. The original reports by Choquette et al. attributed them to mainly thermal effects, claiming that the frequency dependent material gain curve is redshifted by current heating relative to the mean of the two mode frequencies. The relative gain for the two modes is therefore reversed and the originally suppressed mode can start lasing [13]. A more detailed rate-equation model that incorporates four magnetic sublevels [14] was later developed by San Miguel, Feng and Moloney, obtaining its usual name (SFM model) from the creators' initials. For the magnetic sublevel population dynamics, the model considers mechanisms much faster than the thermal response, such as the phase anisotropy γ_p

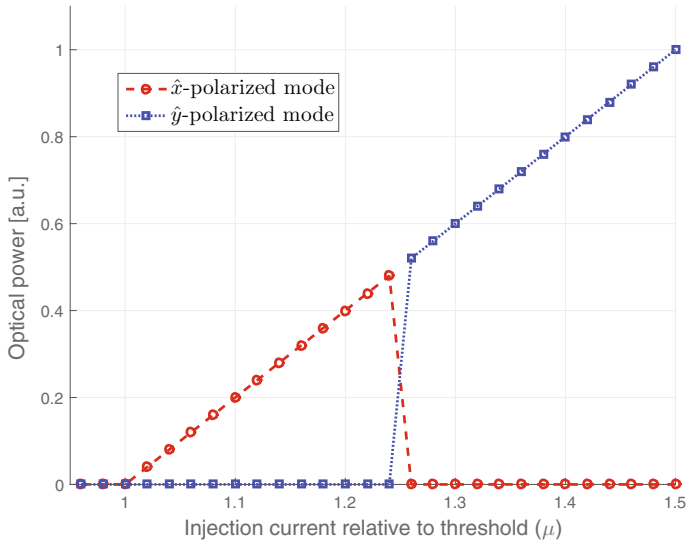


Fig. 4.1 A simple example of polarization switching in a polarization-resolved light-current characteristic. Above threshold ($\mu = 1$), the \hat{x} -polarized mode starts to lase (red circles). Around $\mu = 1.25$ there is an abrupt switch to the \hat{y} -polarized mode (blue squares). The total optical power (the sum of the individual modes) is linearly increasing with current. Based on figures in [12]

(birefringence), the amplitude anisotropy γ_a (the outcome of both gain anisotropies and dichroism), the spin-flip relaxation rate γ_s (mixing of carriers with opposite signs of angular momentum), saturable dispersion α , etc. The validity of the model was shown by an in-depth numerical mode stability analysis conducted by Martín-Regalado et al., making it clear that polarization switching can be explained by setting practical nonzero values to γ_p and α . For a better agreement with experimental data, small but nonvanishing γ_a values are also necessary [12].

Obtaining the critical parameter values, different stability regions can be found in the parameter space. Some of the parameters (γ_a , γ_p , γ_s , α) are built-in, can only be changed on a manufacturing level (or indirectly, if they are current or temperature dependent), but the injection current (occurring in the model as μ , current relative to the threshold) is a controllable external parameter. The article reports regions where only one of the polarization modes (\hat{x} or \hat{y}), both of them, or neither of them are stable. Regions that are bistable correspond to hysteresis cycles in switching (see Fig. 4.2). Switching points are not the same for increasing and decreasing currents, because in bistable regions the already dominating polarization is retained. This agrees well with experimental findings [15].

Quantum key distribution is not the only polarization sensitive application where polarization switches are unwanted. Such are communication links with polarization dependent elements, sensing, optical mice, etc. In these cases, PS cancels the low price advantage of VCSELs, because reliable operation is a more important factor. This called for solutions to mass-produce devices that exhibit stable linear

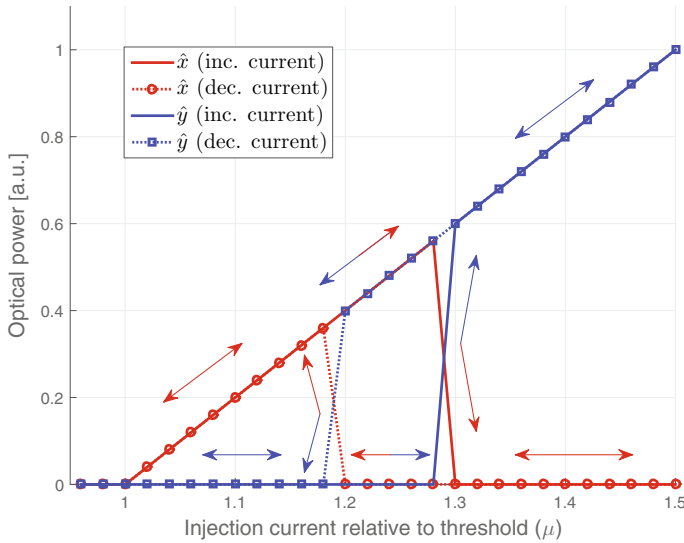


Fig. 4.2 A simple example of polarization switching with hysteresis in a polarization-resolved light-current characteristic. Above threshold ($\mu = 1$), the \hat{x} -polarized mode starts to lase (red). Increasing the current, around $\mu = 1.29$ there is an abrupt switch to the \hat{y} -polarized mode (blue). By decreasing the current, the \hat{y} -polarized mode retains its stability until $\mu = 1.19$, where the polarization is switched back to its original state. We can conclude that the parameter space has a bistable region for current values $1.19 < \mu < 1.29$. Arrows indicate the change in current along the curves. Based on figures in [12]

polarization, like an EEL, but not sacrificing the beneficial properties mentioned earlier. Several different proposals and methods have been tried, including solutions based on asymmetric resonators, external optical feedback, polarization dependent gain or polarization dependent mirrors. In terms of reliability, surface gratings proved to be the best, finding their way into commercial mass production [16].

In the framework of BB84, it is very easy to understand why unwanted polarization switching is problematic in polarization-based QKD protocols. If a photon is sent and measured in the same basis, the raw bit values will be kept; however, a PS is always happening between orthogonal states, therefore all these events will result in a mismatch between Alice's and Bob's raw keys. Two different problems arise from this simple error: Alice and Bob might keep the mismatching bits and use them incorrectly in the final key, or use them during the sifting process. The latter case can potentially lead to false alarms, when the two parties mistakenly think that there is an eavesdropper in the middle.

If switching happens, but the sending and measuring bases are different, the results will be discarded, therefore these switches will not manifest in raw key errors. We can thus conclude that on average, fifty percent of polarization switches will lead to raw key bit errors. Table 4.1 sums up all potential cases in a short example.

Table 4.1 Example of errors in BB84 caused by unwanted polarization switching. States \uparrow and \searrow carry a logical 1, \rightarrow and \nearrow carry a logical 0. Polarization switches occur in bits #2 and #4, but only the latter appears as an error in the raw key due to the measurement basis choices [17]

Bit number	1	2	3	4	5	6
Intended state s_A	\uparrow	\nearrow	\nearrow	\rightarrow	\uparrow	\searrow
Sent state s'_A	\uparrow	\searrow	\nearrow	\uparrow	\uparrow	\searrow
Meas. basis m_B	+	+	\times	+	\times	\times
Alice's raw key s_A	1		0	0		1
Bob's raw key s_B	1		0	1		1

For the B92 protocol, we can also arrive at the conclusion, that on average every second polarization switch will contribute to an error in the raw keys. To analyze the situation, we have to examine three different scenarios:

1. A switch happens, and Bob measures the photon polarization in the **correct basis**. In this case, the measurement outcome will always be $b = 1$, the raw bits $a = a'$ are kept, leading to an error in the raw key.
2. A switch happens, and Bob measures the photon polarization in the **wrong basis**.
 - (a) If the measurement result is $b = 0$, the raw bits a, a' are surely discarded, leading to no error.
 - (b) If the measurement result is $b = 1$, the raw bits $a = \overline{a'}$ are kept, leading to no error.

Therefore, we can conclude that polarization switching only causes problems if Bob chooses the measurement basis correctly, happening in 50% of the cases. The negative effects are the same as mentioned for BB84: false eavesdropping alarms and/or non-matching key bits. Table 4.2 shows an example of how polarization switching can affect the B92 protocol. Every previously mentioned scenario is present: scenario 1 in bit #5, scenario 2a and 2b in bits #2 and #6, respectively.

Table 4.2 Example of errors in B92 caused by unwanted polarization switching. Sending state \rightarrow corresponds to $a = 0$, sending \nearrow corresponds to $a = 1$. Measuring the states \uparrow and \searrow result in a logical 1, measuring \rightarrow and \nearrow result in a logical 0. Measurement basis $+$ corresponds to $a' = 0$, \times corresponds to $a' = 1$. Polarization switches occur in bits #2, #5 and #6, but only #5 appears as an error in the raw key due to the measurement basis choices

Bit number	1	2	3	4	5	6	7	8
Intended state a	\nearrow	\nearrow	\rightarrow	\rightarrow	\rightarrow	\nearrow	\rightarrow	\nearrow
Sent state	\nearrow	\searrow	\rightarrow	\rightarrow	\uparrow	\searrow	\rightarrow	\nearrow
Meas. basis a'	+	+	\times	+	+	+	+	\times
Meas. result b	1	0	0	0	1	1	1	1
Alice's raw key a	1				0	1	0	1
Bob's raw key $\overline{a'}$	1				1	1	0	1

4.3 Polarization Modulation and a Newly Proposed BB84 Transmitter Design

In several applications, polarization switching is seen as a possibility rather than a problem. There is a demand for controlled on-demand polarization switching, most importantly all-optical signal processing (e.g. shift registers [18]). We now propose a new design for the transmitter of the BB84 QKD protocol, deliberately taking advantage of the PS mechanism. As seen before, BB84 uses four different quantum states, forming two pairs of orthogonal bases. This is the main reason why VCSEL polarization switching, happening between two orthogonal modes, may be utilized.

4.3.1 Proposed Design for BB84 Transmitters

To understand why the new design has good prospects, we first describe the so called “trivial transmitter”, depicted in Fig. 4.3. This device contains four individual, linearly polarized laser diodes, either EELs or polarization stabilized VCSELs, each oriented so that their polarizations are aligned to one of the four possible output polarizations used in BB84 [19]. Each selection bit responsible for choosing a transmission basis therefore selects a group of two lasers, while the key bit decides which of the light sources is actually going to be used from within that group. The trivial transmitter is a very simplistic design, with an important advantage that if the lasers are identical, all of them can be driven by identical current pulses.

The new design uses a largely different approach, consisting of only two VCSELs as photon sources, both of them being responsible for emitting two of the four states through polarization modulation. The orientation of the two VCSELs is such that

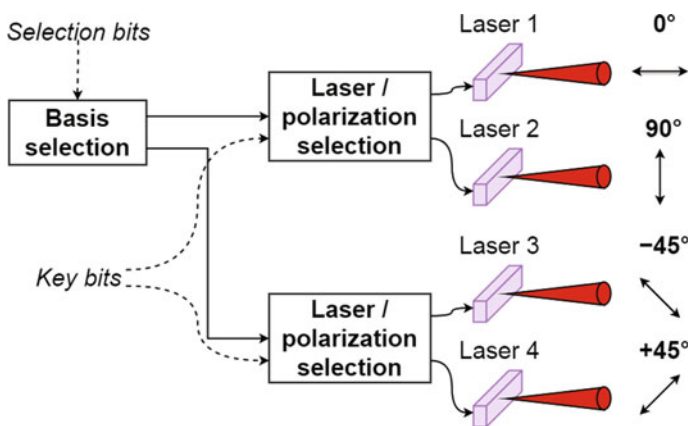


Fig. 4.3 Trivial BB84 transmitter design using four linearly polarized lasers [17]

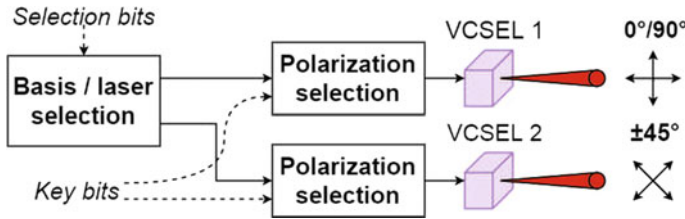


Fig. 4.4 Proposed new BB84 transmitter design using two polarization-switchable VCSELs [17]

their polarization eigenmodes are aligned to the rectilinear and diagonal basis vectors, respectively. In this case, the selection bits not only choose the basis in which a key bit is to be encoded, but simultaneously the laser as well. Key bits are responsible for choosing the desired polarization within the selected basis: they are essentially modulating the laser in polarization. The block diagram of the new transmitter is depicted in Fig. 4.4.

The main benefit of this newly proposed scheme is that it only needs two semiconductor lasers (VCSELs), as opposed to four as seen in the trivial design. This way, costs and the required space can be effectively reduced compared to the trivial BB84 transmitter, at the expense of introducing driving and processing complexities, which are explained in detail in Sect. 4.3.3. As an additional advantage, only two light signals have to be combined instead of four before transmission, reducing the need for a 4:1 coupler to a 2:1 coupler. So far, we treated polarization modulation as a given capability, but the exact means of it should still be investigated.

4.3.2 On-Demand Polarization Switching

There are multiple ways to implement polarization modulation or on-demand polarization switching. One of these is the injection of external linearly polarized light [20]. Although this is a well-documented and prosperous option for applications in all-optical signal processing, the external light is mostly provided by master lasers. This fact is not desirable for the polarization modulated BB84 transmitter, because it cancels one of the original goals, to reduce the number of lasers in the system. We are therefore left to control one of the parameters that play a role in the switching mechanism.

As mentioned in Sect. 4.2.2, most of these parameters cannot be accessed as they are built-in device properties. We can either model them as fixed for a specific VCSEL or slightly current-dependent (specifically γ_p and γ_a) [12], but we cannot rely on them for polarization modulation. The only externally controllable parameter in the model is the pump parameter μ , the injection current normalized to threshold. Although the original reports specified a thermal process, current induced self-heating as the cause of polarization switching, thermal changes do not facilitate fast key transmission rates

due to the large time constants associated with the phenomena. Polarization switches can be expected even when the active region temperature is kept constant [21]. For this reason, we intend to work under thermally stabilized conditions.

Current induced polarization modulation has been studied throughout the last decades, with most of the research focusing on sinusoidal modulation with the VCSEL biased near the switching current. Different types of surface-emitting lasers have been examined. Based on the studies, gain-guided circular VCSELs possess very limited polarization modulation frequencies. One report in 1994 found that the maximum frequency where switching still occurred in a certain device was 80 kHz [22], a 2002 article analyzing another laser measured 90 kHz as the highest frequency where switching occurred in at least 80% of the sine cycles [23]. These slow modulation results can be mainly attributed to thermal processes, that seem to be the dominant cause of PS in gain-guided VCSELs. For index-guided VCSELs, however, only a small percentage of switches might be caused by a thermal origin. As these lasers can be modulated in polarization at much larger frequencies, faster background mechanisms also play an important role in switching [24]. As an example, polarization modulation frequencies up to 50 MHz have already been reported [25].

Lasers in BB84 transmitters do not operate with sinusoidal modulation around a bias point. They are used in pulsed operation, where current is rapidly increased from below threshold to a certain value above threshold, then decreased back. As far as we can tell, no extensive theoretical or practical studies have been conducted to inspect pulsed polarization modulation in VCSELs, but some mentions can be found. One of them (written in 1998) states that the examined VCSEL, which exhibited polarization switching in continuous wave (CW) mode with very slow current changes, emitted stable linearly polarized light under pulsed operation [25]. The current pulses were short, 22 ns long, at a repetition rate of 1 kHz. The authors emphasized the thermal nature of switching in their explanation, stressing that the pulse duration was not enough for current heating to take place. Another brief mention, on the other hand, describes observed polarization switching using current pulses only 10 ns long [24]. This experiment was conducted under two different circumstances, both providing successful results: biasing the laser at threshold and increasing the current, or biasing it well above the DC polarization switching current and using pulses to bring it under the switching point. The last result can be very promising for our application, because it provides a simple solution to modulate the VCSEL in polarization using short current pulses.

Another approach for current-induced polarization modulation is to use VCSELs with special designs. A report from 1994 (Choquette et al.) states that cruciform VCSELs can be polarization modulated with frequencies up to 50 MHz. This is done by applying a small-signal modulation current upon the bias which is set near the DC switching point. In the same article the authors also experimented with large-signal modulation between currents just below the threshold and just above the switching point. At low frequencies, the power in the originally stable linear polarization exhibited a frequency-doubling pulsing. By increasing the frequency, the second pulse per every cycle was reduced in intensity, ultimately disappearing [22].

Asymmetrical current injection techniques may also find their way in the applications. A 2014 study by Barve et al. presented a VCSEL with two asymmetrical sets of electrodes [26]. The article showed that it is possible to modulate the two orthogonal eigenmodes of the laser simultaneously and independently, using one set of electrodes for each mode, with data rates up to 4 Gbps, a huge leap in polarization modulation speed compared to all other current-induced methods. Some of these studies outline exciting solutions for our scheme, presenting ideal VCSEL candidates, but only as long as the cost increase due to the special design requirements is sufficiently low, not to go against our original proposal.

A more in-depth research should be done in the future concerning the pulsed current induced polarization modulation of VCSELs, both theoretically and experimentally. The starting point might be the implementation shown in Fig. 4.5 in case of a single VCSEL. For low currents, \hat{x} -polarization, for high currents, \hat{y} -polarization is dominant, and there is a bistable region between the dashed and dash-dotted lines, where the laser retains its polarization. The small-amplitude pulse on the left is thus perfectly \hat{x} -polarized. To obtain \hat{y} -polarized pulses, we should use a large-amplitude pulse (on the right), the amplitude of which exceeds the higher limit of the bistable region, arriving at a domain where only \hat{y} -polarization is stable. $\mu_p(\tau)$ is denoting

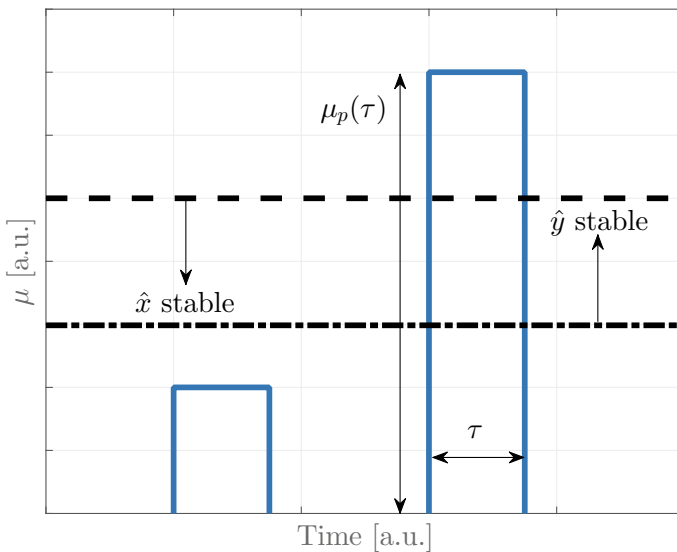


Fig. 4.5 Different amplitude current pulses for producing differently polarized light pulses. The bottom line signifies the threshold current $\mu = 1$. The \hat{x} -polarized mode is stable for currents below the dashed line, the \hat{y} -polarized mode is stable above the dash-dotted line, therefore the low-amplitude pulse will be mainly \hat{x} -polarized, while the high-amplitude pulse should be mainly \hat{y} -polarized. A bistable region, responsible for hysteresis, lies between the two lines. τ is the pulse width, while $\mu_p(\tau)$ is the smallest current peak value needed for polarization switching to occur given a pulse of length τ [17]

the smallest possible current amplitude for which a switching is expected to happen, given a current pulse of length τ .

4.3.3 *Difficulties in the New Design*

The proposed transmitter comes with several problems. In a trivial transmitter, each laser (all assumed to be nearly identical) are driven by identical current pulses, thus the new design introduces additional driving complexity due to the differing current pulses used to modulate the VCSELs in polarization. Different current amplitudes lead to different output powers, but both need to be attenuated to the same final value, otherwise logical zeroes and ones could be distinguished by the power level. Changing the attenuation needs to be fast enough not to be the bottleneck that limits the key transmission rate.

Another problem arises from the fact that pulsed polarization modulation will not produce perfectly polarized pulses in all cases. The pulse on the right in Fig. 4.5 will only be partially \hat{y} -polarized, because the rise and fall times are not infinitely fast, therefore the laser will pass time in the \hat{x} -stable and bistable regions. The lower the \hat{y} -polarized fraction is in the pulse, the higher the probability that the surviving single photon will be erroneously polarized. This has the same effect in the protocol as an unwanted polarization switching would have (Sect. 4.2.2). Because of this, the time evolution of the intensities measured in the orthogonal polarizations should be carefully analyzed. For example, if the pulse start or the pulse end is mainly \hat{x} -polarized, a gating function should be used to block these portions, increasing the correctly polarized fraction in the pulse. To ensure maximum temporal overlap between different quantum states, every pulse should undergo the same pulse shaping, even if they are perfectly polarized, otherwise, time-of-arrival measurements could leak information to an eavesdropper.

The most obvious solution to these difficulties is to use a lossy optical modulator, e.g. an electro-absorption modulator (EAM), that changes its absorption coefficient, therefore its attenuation, as a function of an external voltage. EAMs are available up to data rates around 40 Gbps in classical optical telecommunication links, which makes them a perfect candidate in terms of speed for our design. Single modulators do not usually have high enough total attenuation to reach the single photon power regime, therefore they should be cascaded with a fixed value optical attenuator that provides most of the necessary losses. The modulation depth (>10 dB) provided by the EAMs is also enough to cover the power differences between light pulses generated by two different amplitude current pulses. In addition to standardizing the power levels, EAMs are also suitable for the needed pulse shaping. However, if the price of the EAM exceeds that of two laser diodes, some of the benefits of the new design would be forfeited. Figure 4.6 shows the block diagram of the transmitter complemented with the attenuator and the modulator.

If both of the light pulses are properly polarized and no pulse shaping is necessary, an easier and less costly approach can be followed by putting properly aligned

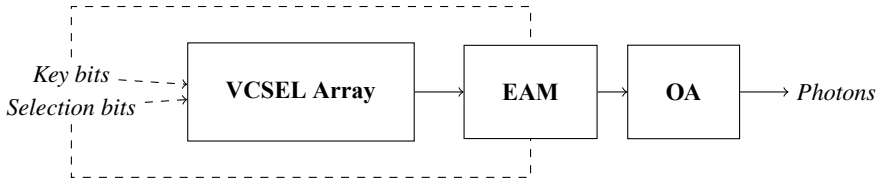


Fig. 4.6 Block diagram of the whole transmitter using a modulator. It includes the VCSEL array depicted in Fig. 4.4, a fixed value optical attenuator (OA) and an electro-absorption modulator (EAM) controlled by the key and selection bits to match the exact attenuation to the desired quantum state

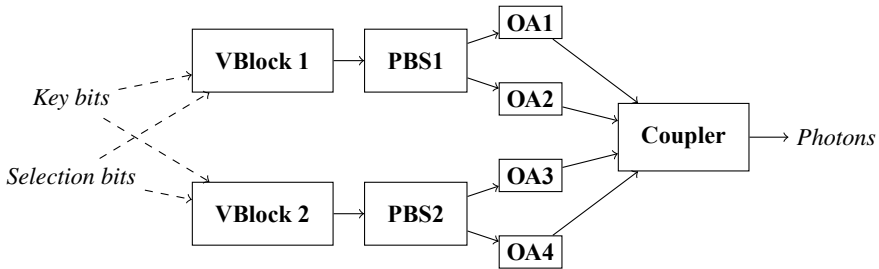


Fig. 4.7 Block diagram of the whole transmitter using a beam splitters. It includes two VCSEL blocks with basis and polarization selection circuitry (VBlocks), two polarization beam splitters (PBS), four fixed value optical attenuators (OA) and a 4:1 optical coupler

polarization beam splitters after the VCSELS to separate pulses with orthogonal polarizations. A fixed value attenuator (or a low speed variable attenuator to counteract long term changes) is placed in each path, carefully calculated individually for each quantum state. The paths are then recombined and the qubits are ready to be transmitted (Fig. 4.7).

The main inconvenience of the proposed design is that switching parameters differ from VCSEL to VCSEL, even for lasers coming from the same wafer. Some of them might not exhibit PS at all. This means, that before installing, every individual diode should be hand-picked and examined to perfectly understand its switching mechanism. Once these are known, the transmitter parameters (attenuation, pulse shaping) need to be individually tuned for every VCSEL.

4.4 Spectral Attacks and Protection

After carefully attenuating and shaping the pulses to ensure that they are as similar in every aspect as we can make them (apart from their polarization), differently polarized photons from a VCSEL will still possess different frequencies due to birefringence. The frequency split between them can be in the order of several tens of GHz [12], which is more than the practical frequency resolution obtainable today.

This way, different quantum states coming from the same VCSEL might be distinguished spectrally. In the trivial design, this hazard can be easily countered by using four lasers with largely overlapping spectra.

If the four different BB84 states have four different frequencies, Eve (who we assume to possess every device that is not forbidden by the laws of physics) only needs to find out the one-to-one correspondence between the frequencies and the polarizations. If this correspondence is unknown at first, Eve would have to guess and choose from the 24 possible permutations after measuring the frequencies and resending some new photons. Choosing wrong would alert Alice and Bob through the induced errors and they could restart that process. This, in turn, is announced publicly on a classical channel, therefore Eve can switch to a different permutation and try again. The correspondence can thus be found in at most 24 turns. After Eve determined how frequency and polarization are related, she can measure the frequency, determine the basis and bit value sent by Alice, then send a quantum state that is identical to the original both in polarization and in frequency. Alice and Bob will never be alarmed, and Eve obtains a perfect third copy of the key.

A practical, but not provably secure solution is to use VCSELs with a very low frequency split that cannot be resolved within experimental accuracy. A provably secure, but not necessarily practical protection against spectral attacks can be achieved by using two identical VCSELs that share the same frequencies f_{LOW} and f_{HIGH} for their two orthogonal polarization eigenmodes (again, the argument of largely overlapping spectra is a more realistic expectation). This way, the one-to-one frequency-polarization correspondence is broken, which in itself is enough to protect the transmission against destructive frequency measurements, where the original photon is annihilated and Eve needs to resend another. Although after the basis choice disclosure she can learn which frequency is assigned to which bit value in a certain basis, a simple frequency measurement will never leak enough information about the polarization.

This solution can be extended even against non-destructive frequency measurements, where photons are not annihilated and their polarization remains unchanged, by a simple frequency-to-bit assignment. If VCSEL 1 has frequency f_{LOW} corresponding to a logical zero and f_{HIGH} to a logical one, the assignment should be the inverse for VCSEL 2 (Table 4.3). In this case, Eve will know the frequencies without errors being registered at Alice's and Bob's, but since there is no one-to-one frequency-bit correspondence, the information gained by the eavesdropper is irrelevant [17]. We have to note here, that non-destructive frequency measurements may not alter the polarization, but still change the photon's wavefunction in other ways. These might alert the two parties, e.g. by increasing the probability that a photon is detected in a wrong time bin, leading to an increasing error rate. It is a consequence of the time-energy uncertainty relation $\Delta E \cdot \Delta t \geq \frac{\hbar}{2}$, and a photon's frequency is related to its energy by $E = h \cdot f$.

Table 4.3 Solution against the spectral attack: Bit value assignment to the low and high frequency modes of identical VCSELs [17]

Mode frequency	Bit value	
	VCSEL 1	VCSEL 2
f_{LOW}	0	1
f_{HIGH}	1	0

4.5 Conclusion

In low-power quantum key distribution protocols, vertical-cavity surface-emitting lasers possess multiple beneficial properties over edge-emitting lasers. There is, however, an inherent problem originating from their polarization switching which may induce errors in practical QKD implementations that use polarization qubits. This calls for the need to use VCSELs in a way that provides control over the polarization of emitted photons. We have shown the nature and frequency of raw key bit errors caused by PS in the BB84 and B92 protocols. As the central point of our work, we outlined a simplified design for the transmitter used in BB84. The new transmitter utilizes polarization switching that occurs between two linear polarization modes orthogonal to each other, requiring only two lasers compared to four in a trivial design. Both of these VCSELs are modulated in polarization. The methods for pulsed mode polarization modulation are discussed. Although specially designed lasers with two sets of electrodes are very promising, to keep the transmitter simple—one of its original assets—, using different amplitude current pulses seems to be the most fitting option to select the desired polarization state. The additional concerns of the new design are also investigated; namely, the need for varying external attenuation due to the varying emitted power caused by differing current amplitudes, and the possibility of a spectral attack performed by eavesdroppers owing to the frequency split between a VCSEL's polarization eigenmodes. The feasibility of our proposal is yet to be validated experimentally, to compare it to already existing solutions in terms of costs, reliability and the maximally achievable key rate.

Acknowledgements The authors would like to thank Dr. Zsolt Kis (Wigner Research Centre for Physics, Budapest, Hungary) for the helpful suggestions regarding the physics and background of photon frequency measurements.

References

1. C.E. Shannon, Communication theory of secrecy systems. *Bell Labs Tech. J.* **28**, 656–715 (1949)
2. W.K. Wootters, W.H. Zurek, A single quantum cannot be cloned. *Nature* **299**, 802–803 (1982)
3. C.H. Bennett, G. Brassard, Quantum cryptography: public key distribution and coin tossing, in *Proceedings of IEEE International Conference on Computers, Systems and Signal Processing*, vol. 175 (IEEE, 1984). p. 8

4. C.H. Bennett, Quantum cryptography using any two nonorthogonal states. *Phys. Rev. Lett.* **68**, 3121 (1992)
5. A.K. Ekert, Quantum cryptography based on Bell's theorem. *Phys. Rev. Lett.* **67**, 661 (1991)
6. C.H. Bennett, F. Bessette, G. Brassard, L. Salvail, J. Smolin, Experimental quantum cryptography. *J. Cryptol.* **5**, 3–28 (1992)
7. R.J. Glauber, Coherent and incoherent states of the radiation field. *Phys. Rev.* **131**, 2766 (1963)
8. G. Brassard, N. Lütkenhaus, T. Mor, B.C. Sanders, Limitations on practical quantum cryptography. *Phys. Rev. Lett.* **85**, 1330 (2000)
9. R. Michalzik (ed.), *VCSEL Fundamentals* (Springer, Berlin, 2013) pp. 19–75
10. G. Vest, M. Rau, L. Fuchs, G. Corrielli, H. Weier, S. Nauerth, A. Crespi, R. Osellame, H. Weinfurter, Design and evaluation of a handheld quantum key distribution sender module. *IEEE J. Sel. Top. Quantum Electron.* **21**, 131–137 (2015)
11. S. Nazhan, Z. Ghassemlooy, Polarization switching dependence of vcsel on variable polarization optical feedback. *IEEE J. Quantum Electron.* **53**, 1–7 (2017)
12. J. Martín-Regalado, F. Prati, M. San Miguel, N. Abraham, Polarization properties of vertical-cavity surface-emitting lasers. *IEEE J. Quantum Electron.* **33**, 765–783 (1997)
13. K.D. Choquette, R.P. Schneider, K.L. Lear, R.E. Leibenguth, Gain-dependent polarization properties of vertical-cavity lasers. *IEEE J. Sel. Top. Quantum Electron.* **1**, 661–666 (1995)
14. M. San Miguel, Q. Feng, J.V. Moloney, Light-polarization dynamics in surface-emitting semiconductor lasers. *Phys. Rev. A* **52**, 1728 (1995)
15. A.B. Kaplan, Investigating the polarization properties of vertical-cavity surface-emitting lasers. B.A. honors thesis, Amherst College (2007)
16. R. Michalzik, J.M. Ostermann, Polarization control of VCSELs, in R. Michalzik (ed.) *VCSELs*. (Springer, Berlin, 2013) pp. 147–179
17. Á. Schranz, E. Udvary, Transmitter design proposal for the BB84 quantum key distribution protocol using polarization modulated vertical cavity surface-emitting lasers, in *Proceedings of the 6th International Conference on Photonics, Optics and Laser Technology* (2018), pp. 252–258
18. T. Katayama, D. Hayashi, H. Kawaguchi, All-optical shift register using polarization bistable VCSEL array. *IEEE Photonics Technol. Lett.* **28**, 2062–2065 (2016)
19. A. Ruiz-Alba, D. Calvo, V. Garcia-Muñoz, A. Martinez, W. Amaya, J. Roza, J. Mora, J. Capmany, Practical quantum key distribution based on the BB84 protocol. *Waves* **3**, 4–14 (2011)
20. S. Bandyopadhyay, Y. Hong, P. Spencer, K. Shore, Vcsel polarization control by optical injection. *J. Light. Technol.* **21**, 2395–2404 (2003)
21. J. Martín-Regalado, J. Chilla, J. Rocca, P. Brusenbach, Polarization switching in vertical-cavity surface emitting lasers observed at constant active region temperature. *Appl. Phys. Lett.* **70**, 3350–3352 (1997)
22. K.D. Choquette, K. Lear, R. Leibenguth, M. Asom, Polarization modulation of cruciform vertical-cavity laser diodes. *Appl. Phys. Lett.* **64**, 2767–2769 (1994)
23. G. Verschaffelt, J. Albert, I. Veretennicoff, J. Danckaert, S. Barbay, G. Giacomelli, F. Marin, Frequency response of current-driven polarization modulation in vertical-cavity surface-emitting lasers. *Appl. Phys. Lett.* **80**, 2248–2250 (2002)
24. G. Verschaffelt, J. Albert, B. Nagler, M. Peeters, J. Danckaert, S. Barbay, G. Giacomelli, F. Marin, Frequency response of polarization switching in vertical-cavity surface-emitting lasers. *IEEE J. Quantum Electron.* **39**, 1177–1186 (2003)
25. K. Panajotov, B. Ryvkin, J. Danckaert, M. Peeters, H. Thienpont, I. Veretennicoff, Polarization switching in vcsel's due to thermal lensing. *IEEE Photonics Technol. Lett.* **10**, 6–8 (1998)
26. A.V. Barve, A. Mehta, A. Husain, L. Coldren, Ultrafast electrical polarization modulation in VCSEL with asymmetric current injection, in *Optical Interconnects Conference, 2014 IEEE* (IEEE, 2014), pp. 91–92

Chapter 5

Tm Based Solid-State Lasers—Toward High Power Tunability—A Review



Salman Noach, Uzziel Sheintop and Rotem Nahear

Abstract Tm based lasers have gained a lot of interest during recent years for many application. Different works have been done applying the Tm Characteristics with different hosts in order to improve the laser performance in many features (power, tunability range, spectral bandwidth, energy per pulse, etc.) by many techniques. In this review we will focused on achieving tunability in Tm solid-state lasers, and we will give an overview of the works have been done in this field.

5.1 Introduction

In this chapter we want to focus on tunable Tm based lasers and to give comprehensive review about different methods and results that have been published since the reviews of Sorokina and Vodopyanov [1], Godard [2], Scholle et al. [3] about Tm lasers.

Tm is one of the most popular active dopants for lasers, having a broad Fluorescence range around 2 μm . This “eye-safe” spectral segment is of growing interest for many applications: microsurgery, material processing, gas monitoring, remote sensing, and as a pump source for MIR lasers. Improvement and commercialization of AlGaAs laser diodes that are used to pump Tm doped lasers at ~ 790 nm contribute to rapidly improving the performance of those lasers in recent years. Because of the high brightness of the pump diode, end pump configuration is the preferred method of pumping these crystals since it provides efficient and high beam quality output.

We will begin with a brief review about the Tm ion and its properties with an emphasis on those properties which contribute to the increase of tunability range. Then we will review the most common methods to achieve tunability: Prism, Grating, BF plate, Lyot filter, Etalons, VBG. Each method has advantages and disadvantages concerning losses, tunability range, damage threshold, and cost. During the review,

S. Noach (✉) · U. Sheintop · R. Nahear
Department of Applied Physics, Electro-Optics Engineering Faculty, Jerusalem College of
Technology, Jerusalem, Israel
e-mail: salman@jct.ac.il

we will remark works that used the different techniques. Most of the works were implemented for CW tunable lasers; a separate section will be dedicated to tunable pulsed lasers.

5.2 Tm Laser Characterizes

5.2.1 Tm^{3+} Ion Spectroscopy

It is accustomed to describe the rare earth electronic configuration on the basis of the xenon one, since the electronic arrangement of the ground state in rare earth atom has a core with the same electronic filling as xenon [4, 5]. The three first shells are entirely occupied, as well as the 4s, 4p, 4d, 5s, and 5p subshells, while the f subshell in the fourth shell, which adapted to contain up to 14 electrons, is empty in xenon. For rare earth ions, the 6s subshell is also completely filled, and the 4f subshell has a different occupation to each element.

The Thulium atom belongs to the lanthanides series of chemical elements with atomic number 69. Its electronic configuration is $[Xe]4f^{13}6s^2$. As in the other Lanthanides, it has a trivalent ionization state that removes the two 6s electrons and one 4f electron in favor of the ionic binding to the host. The ionized Tm^{3+} ion has thus 66 electrons in its +3 oxidation state, with an electronic configuration $[Xe]4f^{12}$. The 4f orbitals are partially filled, and the optical transitions take place within these subshells. The fluorescence spectra of all rare earth ions arise from electronic transitions between levels of the partially filled 4f shell. Electrons present in the 4f shell can be raised by light absorption into unoccupied 4f levels. In fact, the spatial extension of the 5p and 5s orbitals are broader than the spatial extension of the 4f orbital although their energy is comparatively lower. The high energy lying 4f levels are thus shielded by the less energetic 5s and 5p filled states. This shielding effect protects the optically active 4f states from external electrostatic perturbations and from the strong coupling to the host matrix crystal fields.

The effect of confinement of the 4f orbital is called the Lanthanide contraction. As a result, emission lines are relatively narrow, and the energy level structure varies only slightly from one host to another. The effect of the crystal field is usually treated as a perturbation on the free-ion levels. The perturbation is small compared to the split-ups caused by electrostatic Coulomb interactions ($\sim 10,000\text{ cm}^{-1}$ split) and spin-orbit ($\sim 3,000\text{ cm}^{-1}$ split) among the 4f electrons. The primary change in the energy levels is a splitting of each of the free-ion levels in many closely spaced levels caused by the Stark effect of the crystal field, which changes from one host to another. In crystals, the free-ion levels are then referred to as manifolds. In the 4f shell, the electric dipole transitions are parity forbidden, which are enabled by the crystal-field interaction. The low probability of the “weakly allowed” transitions leads to extremely long upper laser level lifetimes (few ms), resulting in significant energy storage. In the splitting hierarchy of the thulium ion’s 4f energy level, the

first tier of splitting is the result of interactions between 4f electrons based on their spin and orbital angular momentum quantum numbers (typically denoted as S and L, respectively). The second split is caused by spin-orbit coupling and is determined by the total angular momentum, J. The final and weakest tier comes from the interaction between 4f electrons and the lattice structure.

In a real Tm doped crystal, the 3F_4 manifold is lying between the 3H_6 and 3H_5 manifolds. An energy level diagram for the laser crystal Tm:YAG [3] is presented in Fig. 5.1.

In the Tm^{3+} ion, the 2 μm laser transition is between the 3F_4 and 3H_6 energy manifolds. The wavelength for the pump laser is typically between 785–795 nm and is readily available from AlGaAs pump diodes. This pump excites the ground state ion into the 3H_4 manifold, where it decays to the 3H_5 manifold via radiative and non-radiative processes, then down to the metastable 3F_4 upper laser level non-radiatively. This transition, which results in a photon with a wavelength of 2.019 μm in YAG, has a lifetime of 10 ms and an emission cross-section on the order of $5 \times 10^{-21} \text{ cm}^2$. The probability of this transition and its emission can either be assisted by cross-relaxation of the pump level, or impeded by upconversion. The Figure also shows the Stark splitting of the ground state, which is important for the thermal population of the lower laser level of the 2 μm laser transition. At room temperature, the higher sublevel of the ground state manifold is populated with about 2% of the ground state total population. This makes the Tm ion a quasi-three-level laser system, as it is sufficient to create inversion population only relative to this higher sublevel. The upper laser level 3F_4 can also be populated by a cross-relaxation process that

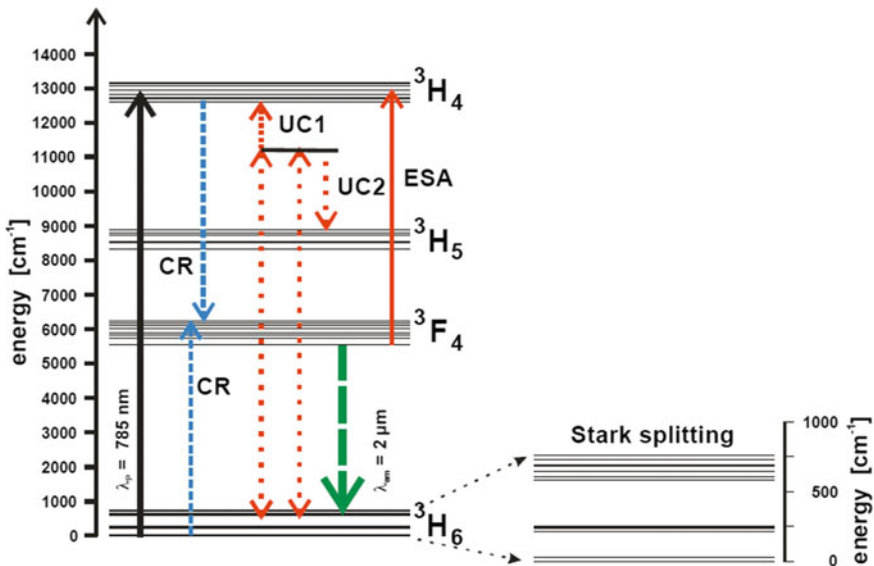


Fig. 5.1 Tm:YAG laser crystal energy levels [3]

occurs between two thulium ions. In this non-radiative process for one ion an electron relaxes from the 3H_4 level to the 3F_4 level, whereas, for an adjacent ion, an electron is excited from the 3H_6 ground state to the 3F_4 level. This excitation process yields two excited ions in the upper laser level for each absorbed pump photon. Therefore, the quantum efficiency is nearly two when the cross-relaxation process is highly efficient. Thus, instead of maximum efficiency of $\sim 40\%$, one can obtain a double efficiency of $\sim 80\%$, in theory. The efficiency of the cross-relaxation process increases with the doping concentration of the thulium ions, since the involved dipole-dipole interaction depends on the Tm adjacent spacing. However, when raising doping concentration, it is essential to take into account other unwanted issues like transition processes e.g. upconversion and pump thermal fracture limit [6].

5.2.2 *Tm Host Materials Properties*

Even though the Tm ion is shielded from the lattice, the host matrix has significant effect on the laser performance. The thermo-mechanical properties of the host determent the scalability toward high power pump. Also, they affect beam quality through thermal lensing and birefringent. The irradiative decay time corresponds to the phonon energy of the host, significantly affect the metastable level lifetime. The lattice symmetry is reflected in the polarization dependent cross section, and the exact emission spectrum is dependent on the lattice field at the ion location. There are some families of materials usually used as hosts for lasers, like, granets, vanadates, etc. [4]. In the following paragraphs, the main properties of those hosts' families will be briefly presented.

The most known garnet host is the YAG. It is known for the excellent thermo-mechanical properties and is widely used with Nd ion. Tm:YAG is also used lasing at $2\ \mu\text{m}$, but is less favorable as the isotropic nature of the YAG also makes it vulnerable to thermal birefringent. Another garnet host is the LuAG. Its thermo-mechanical properties are similar to the YAG. The Lu radii is closer to the lanthanides radii, which enable high doping concentration.

The Tm-doped vanadates are favorable for their high cross sections and broad linewidths compared to YAG, and their thermal conductivity is similar. Another useful property is their naturally polarized emission, neglecting the thermal birefringent effect. The most used vanadates are the Tm:YVO₄ and the Tm:GdVO₄.

The fluorides hosts are softer than the oxides and have smaller thermal conductivity. On the other hand, their emission and absorption cross section are higher, and the lower phonon energy cause very long lifetime beneficial for Q-switch operation. The most common is the YLF and potentially the LiLuF [7].

Another remarkable host is the YAP. Its thermo-mechanical properties comparable with the YAG and higher cross section. The YAP is an anisotropic crystal allowing natural polarization of the laser light.

If the crystal electric field is varied at the dopant position, each ion will have a slightly different spectrum. This will result in larger emission spectrum, enabling

larger tunability and shorter mode lock pulses. In LuYAG some of the ytterbium ions are replaced with lutetium. The thulium ions at different locations have different neighboring ion. In the Tm:CaYAlO₄ structure, the Ca²⁺ and Y³⁺ ions are distributed almost statistically in the nine coordinated sites, and the smaller Al³⁺ ions occupy the position with octahedral symmetry. The doped Tm³⁺ ions can substitute in the Al- and Y-sites. Multi-site replacement and the intense effect of the crystal lattice field together lead to a considerable inhomogeneous broadening of the optical spectra of Tm:CYA [8].

Disordered crystals are characterized by a broad fluorescence linewidth arising from their disordered lattice structures, and thus they could potentially generate shorter mode-locked pulses. Among disordered crystals, calcium niobium gallium garnet (Ca₃Nb_{1.6875}Ga_{3.1875}O₁₂, CNGG) and calcium lithium niobium gallium garnet (Ca₃Li_{0.275}Nb_{1.775}Ga_{2.950}O₁₂, CLNGG) are representative crystals that have excellent optical and thermal properties. The CNGG crystal has a disordered structure due to the random distribution of Nb⁵⁺, Ga³⁺, and cationic vacancies in the host lattices. After introducing Li⁺ into CNGG, the composition ratio of Nb⁵⁺ and Ga³⁺ in the host is changed, causing a different crystal lattice field. The disordered crystal lattice field in the host will cause spectrum inhomogeneous broadening of laser media [9].

5.3 Tunability Methods and Results

5.3.1 Prism Tuning

Prism tuning is based on the material chromatic dispersion. The prism material should be transparent in the region of interest in order to eliminate loss and is quite easy to implement. If the laser source is linearly polarized, applying the prism at the Brewster angle of incidence cause the reflection losses stay very small and one can eliminate the prism surface coating. Prism-tuning method does not have intrinsic bandwidth limitation, but the wavelength selectivity of prism tuning is rather low [10].

A work of achieving tunability using a CaF₂ prism was published by Qin et al. [11] using a Tm:CaYAlO₄ (Tm:CYA) crystal.

Despite the excellent properties of the Tm:CYA crystal [8], the first publication for a laser performance using this crystal was published In 1997 for a continuous wave (CW) laser operation of a 2 at.% Tm³⁺-doped crystal pumped by a Ti:sapphire, in which the laser output power and slope efficiency were only 50 mW and 18% [12]. In 2013 Qin et al. [11] present significant improvement by using a 6 at.% Tm:CaYAlO₄ (Tm:CYA) crystal pumped by a fiber-coupled laser diode around 790 nm. The laser emits a maximum output power of 4.3 W with a slope efficiency as high as 46.7%.

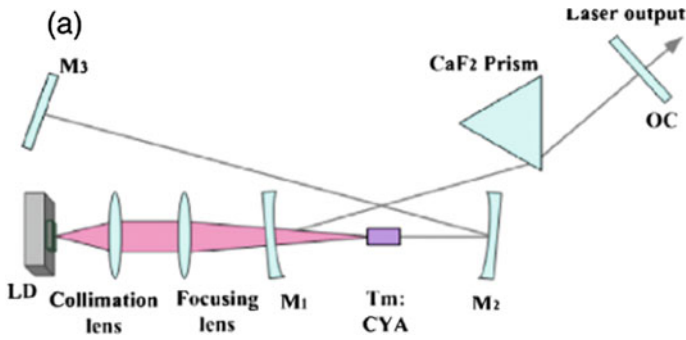


Fig. 5.2 The schematic of the wavelength-tunable Tm:CYA laser [11]

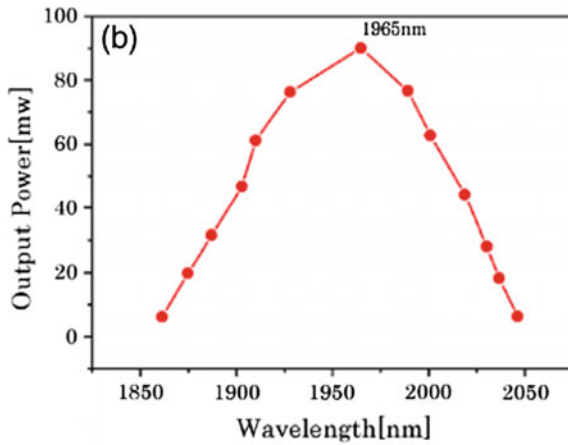


Fig. 5.3 The wavelength tuning trace for the Tm:CYA laser [9]

X-folded cavity setup with an output coupling of 5% is employed, as shown in Fig. 5.2. The pump source is a single-emitter laser diode at 790 nm. A CaF₂ prism with Brewster-angle incidence was used as a wavelength tuning element. By rotating the angle of the output coupler, the output laser could be continuously tuned. Figure 5.3 shows the wavelength tuning curve measured at an incident pump power of 1.69 W. The maximum output power of 90 mW was obtained at 1965 nm. The laser continuously tuned from 1861 to 2046 nm, giving a tuning range of 185 nm and a FWHM bandwidth of 110 nm.

Other work using a prism to achieved tunability published by Gao et al. [13] using Tm:CLNGG disordered crystal. The laser emits a maximum power of 1.17 W with a slope efficiency as high as 42%.

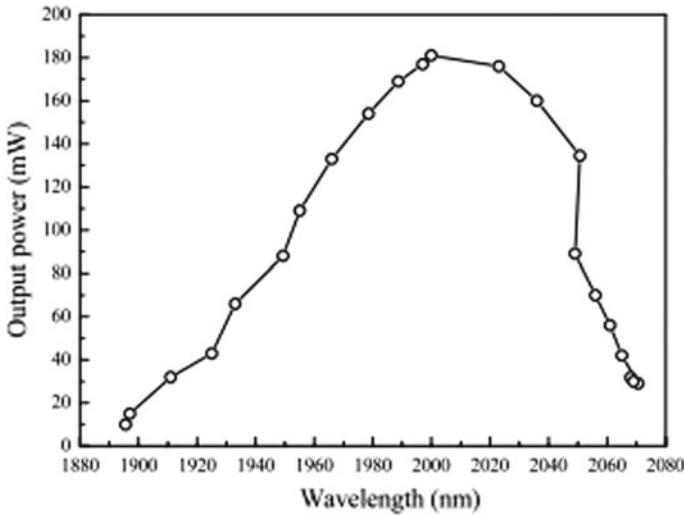


Fig. 5.4 Wavelength tuning trace of the Tm:CLNGG laser [13]

X-folded cavity with an output coupling of 0.5% was employed. A CaF_2 prism was used as a wavelength tuning element in the cavity. To eliminate the Fresnel loss of the crystal surfaces, the Tm:CLNGG crystal was placed at the Brewster angle in the wavelength tuning experiment. A single-emitter LD was used as the pump source. The output laser wavelength could be tuned by rotating the angle of the output coupler. The tuning spectra and corresponding output power of the Tm:CLNGG laser were recorded under an incident pump power of 2.94 W; the tuning trace is shown in Fig. 5.4. The laser wavelength is continuously tunable by changing the angle of the output coupler. A maximum output power of 181 mW was obtained at 2000 nm. A widely tunable range from 1896 to 2069 nm has been achieved, showing that the Tm:CLNGG disordered crystal is very promising for femtosecond pulse generation by mode-locking.

5.3.2 Grating Tuning

Grating tuning is based upon the chromatic dispersion of the diffracted beam. In the Littrow configuration, one of the resonator mirrors is replaced by the grating. The first order beam is reflected at the same angle as the incidence beam, stabilizing the resonator. The condition for the angle matching is given by:

$$\sin \theta = \lambda/2d \quad (5.1)$$

where θ is the angle of incidence on the grating, λ is the wavelength and d is the spacing of the grating, thus allowing only specific wavelength to lase. By tuning

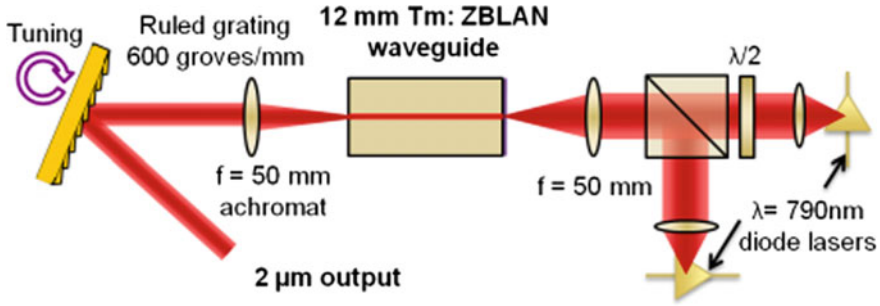
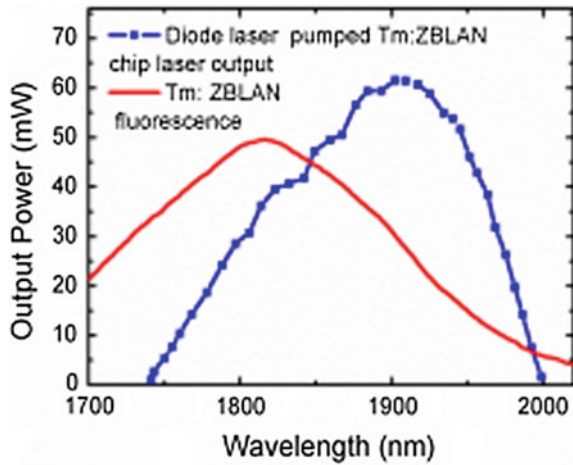


Fig. 5.5 Schematic of the Tm:ZBLAN waveguide laser [14]

Fig. 5.6 Measured tuning range of the Littrow configured Tm:ZBLAN laser for an incident pump power of 475 mW [14]



the incident beam angle, e.g. rotating the grating, the condition applies for different wavelength.

The grating selectivity is about an order of magnitude higher than a prism.

The losses in grating are due to imperfect reflectivity and the non-diffracted beam. For mid IR, a metal coated grating can have a quite high reflectivity, and by replacing the OC by the grating, the non-diffracted beam can be used as the output coupler.

Lancaster et al. [14] published a widely tunable short-infrared thulium and holmium doped fluorozirconate waveguide chip lasers using a grating.

The Tm^{3+} and Ho^{3+} doped fluorozirconate (ZBLAN) glass waveguide extended cavity lasers was pumped by a 790 nm diode by using the grating at the Littrow configuration, the laser giving tunes from 1750 to 1998 nm at a 14% incident slope efficiency, and a beam quality of $1.2 M^2$, tunability max power was 60 mW for pump power of 474 mw.

The laser scheme is shown in Fig. 5.5 and typical spectrum of the tuned output is shown in Fig. 5.6.

5.3.3 Volume Bragg Gratings Tuning

Volume Bragg Gratings (VBG) offers an alternative approach for tuning and line narrowing with reduced losses.

Volume Bragg gratings are Bragg gratings which are written inside some transparent material, e.g. in the form of a cube or a parallelepiped—in contrast to diffraction gratings made on the surface of an optical element or fiber Bragg gratings, where the grating is written into the core of an optical fiber. These gratings are holographic gratings recorded in photo-thermo-refractive (PTR) glass with high diffraction efficiencies (~99%), low insertion loss. PTR glasses are transparent from the UV to the MIR permitting working in many spectral regimes. With a damage threshold of ~40 J/cm² (for 8 ns pulses) and tolerance to continuous wave (CW) laser radiation in the near IR region at least up to several tens of kilowatts per square centimeter [15, 16], a nonlinear refractive index similar to fused silica, and multi-kW (CW) power loading capability VBG are ideal for high intra-cavity powers [17]. The VBG has a narrow spectral line width, the reflectivity bandwidth can be designed from 40 to 1000 pm. Spectral tuning as a cavity reflector requires satisfying the Bragg condition for angles close to normal incidence, which is expressed by the equation:

$$2nd \cos \theta = m\lambda \quad (5.2)$$

where n is the index of refraction of the glass, d is the spacing of the grating planes, θ is the angle of incidence on the grating, m is the diffraction order, and λ is the wavelength. Limited spectral tuning can also be obtained by temperature tuning, where the diffraction wavelength changes by 7 pm/K.

The VBG serves as a selective wavelength mirrors, and usually placed in a laser resonator which enables spectral and thermal management of the laser radiation. Its characteristics facilitate construction of compact tunable solid-state lasers with low cavity loss and high output power. These properties make VBGs very attractive for tunable lasers, especially for Q-switching lasers with high peak power.

Compared with other tuning methods, its cost is still considered expensive.

We can mention some works achieved tunability with VBG, Sun et al. [18] achieved continuously tuning range of ~60 nm from 1935.3 to 1994.9 nm in a Tm:LuYAG laser. Maximum output power of 1.76 W at 1999.7 nm was obtained for a launched pump power of 10.75 W with the VBG used at normal incidence. Laser tunability results achieved by using the four-mirror folded resonator at a constant input power of 6.21 W as can be seen in Fig. 5.7.

The tuning reached by rotates both the VBG and the OC respectively. Extent tuning of the reflected wavelength from the Bragg grating could be achieved. The lasing wavelength λ changes according to $\lambda = \lambda_0 \cos \theta$, where λ_0 is the wavelength at a normal incidence and θ is the internal angle in the VBG. Tuning results are shown in Fig. 5.8.

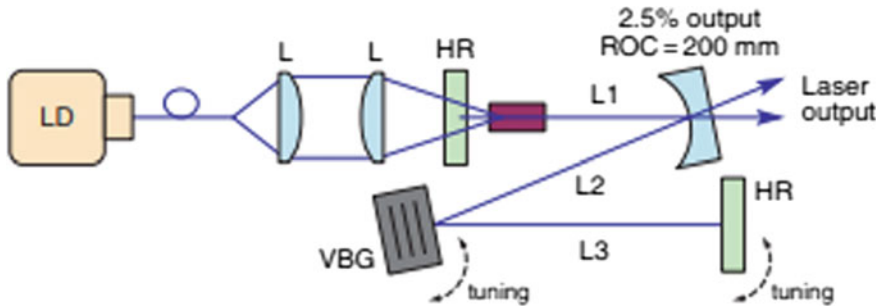


Fig. 5.7 Cavity configuration of the diode-pumped Tm:LuYAG laser: L1 = 129 mm, L2 = 90 mm, and L3 = 90 mm [18]

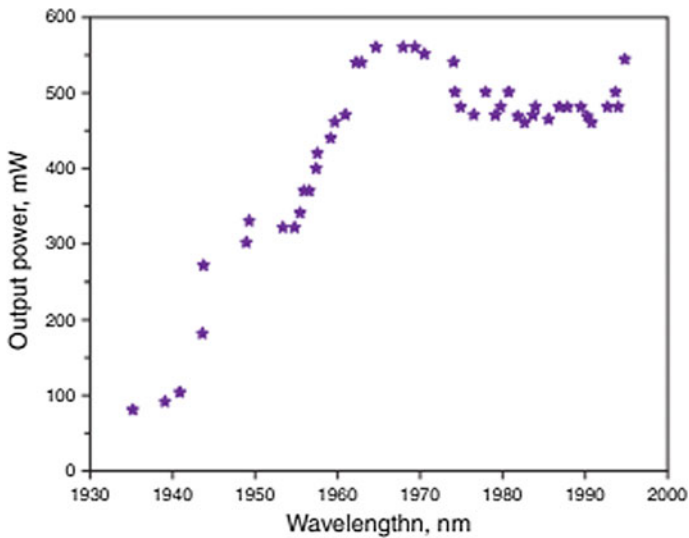


Fig. 5.8 Tunable operation of the Tm:LuYAG laser at 6.21 W incident pump power [18]

Recently a high power Tm:CaGdAlO₄ was published [19]. The laser was pump by Raman shifted Er, Yb laser emitting at 1693 nm. The laser was tune using VBG and maximum output power of 5.8 W was achieved. Tuning rang of 74 nm from 1920 to 1994 nm was demonstrated. In addition linewidth of less then 50 pm was shown Fig. 5.9.

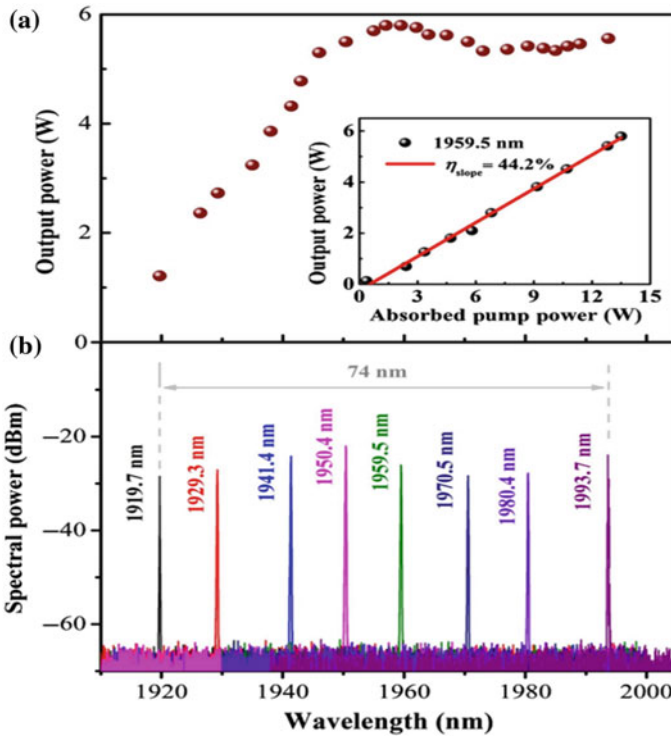


Fig. 5.9 **a** Tunability curve of the Tm:CALGO laser at 13.5 W absorbed pump power, inset is output power at 1959.5 nm versus absorbed pump power. **b** Laser spectra of tunable Tm:CALGO laser [19]

5.3.4 Birefringent Filters Tuning

Birefringent filters make use of the fact that the phase retardation of a birefringent plate is wavelength-dependent [20–22]. Inserting polarization-sensitive losses into the cavity will cause wavelength dependent losses. Due to the periodic nature of the phase there is intrinsic FSR limiting the tuning range.

For normally incident beam the phase retardation is given by:

$$\Gamma = \frac{2\pi}{\lambda} |n_o - n_e| d \tag{5.3}$$

Here, d is the plate thickness, n_o , n_e are the ordinary and extraordinary refractive indices, respectively. If the plate is placed between two parallel polarizers, and the optical axis is 45° with respect to the polarizer axis, the transmission of the electric field is sinusoidal with the wavenumber. The FSR of birefringent plate is

$$\Delta\lambda_{\text{FSR}} \cong \frac{\lambda^2}{d * |n_o - n_e|} \quad (5.4)$$

The FWHM bandwidth of the plate is

$$\Delta\lambda \cong \frac{\lambda^2}{2 * d * |n_o - n_e|} = \frac{\Delta\lambda_{\text{FSR}}}{2} \quad (5.5)$$

It is evident that increasing the free spectral range will increase the bandwidth but will decrease the wavelength selectivity. In Lyot configuration a set of birefringent plates is used, each of them is placed between two polarizers. Each of the plates is twice the width of the previous one. In this filter the free spectral range is set by the thinnest plate while the FWHM bandwidth is set by the thickest one, allowing large FSR and good selectivity.

Because of the dependence of n_e in the incident beam polarization, changing the angle of the optic axis will tune the transmitted wavelength.

When used inside laser cavity for tuning, the need of high selectivity is low due to the laser own selectivity. On the other hand inserting number of polarizers and coating the plate is expensive. For these reasons Lyot inside laser cavity is composed from birefringent plates placed in Brewster angle. At this angle, the plates do not need coating and create polarization dependent losses. A complete treatment of the Lyot filter inside the cavity can be found [23], but the main aspect of the FSR and bandwidth dependent in the plate thickness and the benefits of using few plates are the same.

The simplicity of inserting the birefringent plate into the cavity and the large FSR, low cost and low losses made the birefringent plate to be very common in tunable lasers. The birefringent plate force the laser to specific polarization even when its gain smaller e.g. Tm:YLF laser with the birefringent plate optimize for 1910 nm, when tune to 1885 will have far less gain in the polarization chosen than in the orthogonal one.

We shall mention some of the laser schemes and tunability results. Ereimeikin et al. [24] published a paper about investigation the effect of water vapor on the laser spectrum in Tm:YLF. The free running laser spectrum was changing with different pump power. The laser wavelengths coincide with the water vapor absorption line causing a broad spectrum. The laser scheme, with using two diodes from both sides of the gain medium in order to increase the pump efficiency and filtering the pump wavelength, is shown in Fig. 5.10. The laser resonator was placed inside a nitrogen chamber to control humidity.

Inserting 5 mm sapphire plate at Brewster angle allows tuning of more than 70 nm (1860–1940 nm), shown in Fig. 5.11, with max power more than 2.5 W a decrease of less than 1.5% respect to free running operation. When the laser was tuned from the water vapor absorption lines the spectrum was smaller than 1 nm. It is evident there is no peak output power around 1880 despite the π polarization has significant peak in the emission cross section, due to the birefringent plate preferring the σ polarization.

Fig. 5.10 Scheme of the experimental setup: (1) highly reflecting mirror; (2) output mirror; (3) Tm:YLF crystal; (4) fiber pigtailed of the pump diode lasers; (5) dichroic mirrors; (6) optical condensers; (7) phase plate; (8) semitransparent mirror; (9) power meter; (10) MDR-41 spectrograph; (11) FSA-G1 photodetector; (12) nitrogen-pumped chamber [24]

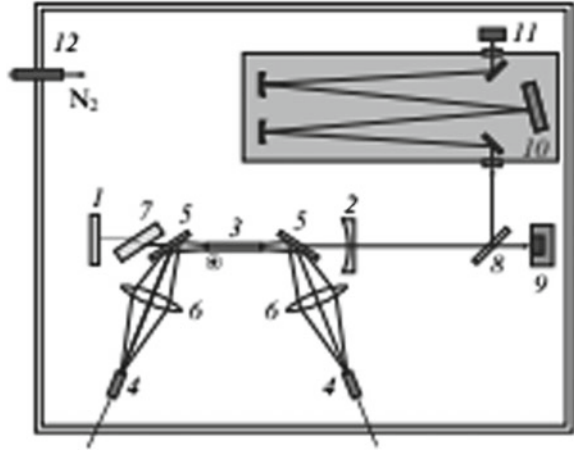
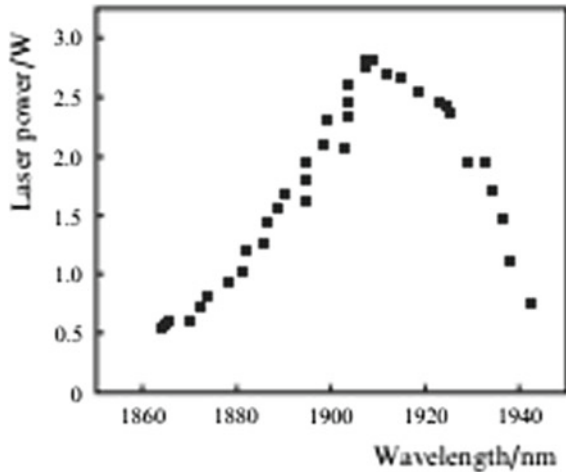


Fig. 5.11 Tuning curve of the Tm:YLF laser with a phase plate at the pump power of 18 W [24]



Another high power tunable laser, based on Tm:GdVO₄, published by Yu-Feng et al. [25]. The Tm-doped vanadates are favorable for their high cross sections and broad linewidths compared to YAG, and their thermal conductivity is similar. Another useful property is their naturally polarized emission, neglecting the thermal birefringent effect. A tuning range of 130 nm with max power of 2.5 W was achieved. The laser scheme, pumping the crystal from both sides in order to decrease the thermal effect, is shown in Fig. 5.12, while the laser tunability is shown in Fig. 5.13.

During 2007 Coluccelli et al. [26] published a letter about the laser performance of Tm:LiLuF. The tunability range obtains with Lyot filter was 230 nm with max power of 0.7 W. the laser scheme is shown in Fig. 5.14, and the laser tunability is shown in Fig. 5.15.

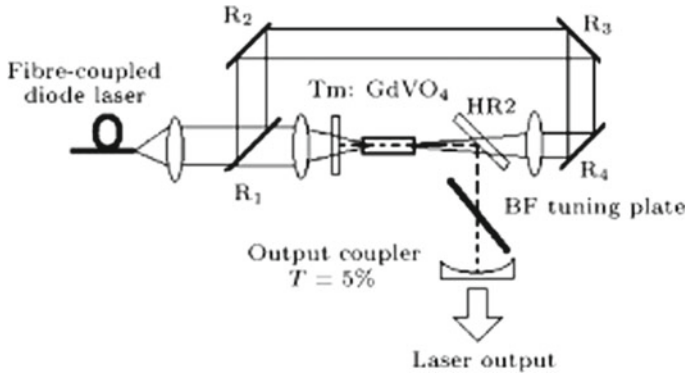


Fig. 5.12 Cavity configuration of the diode pumped Tm:GdVO₄ laser [25]

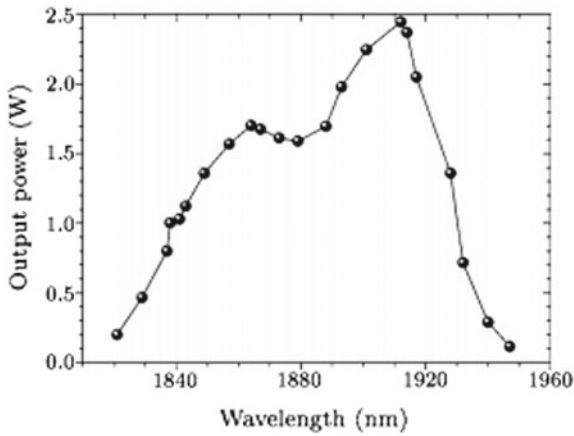


Fig. 5.13 Tunability curve of the Tm:GdVO₄ laser [25]

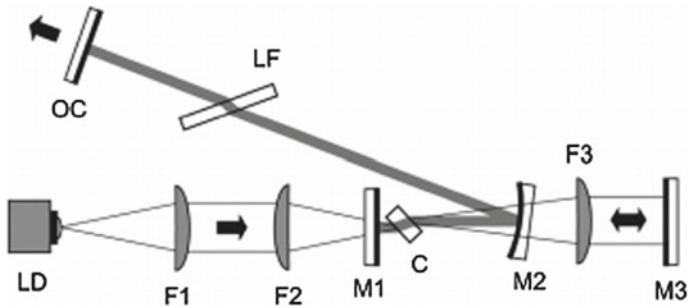
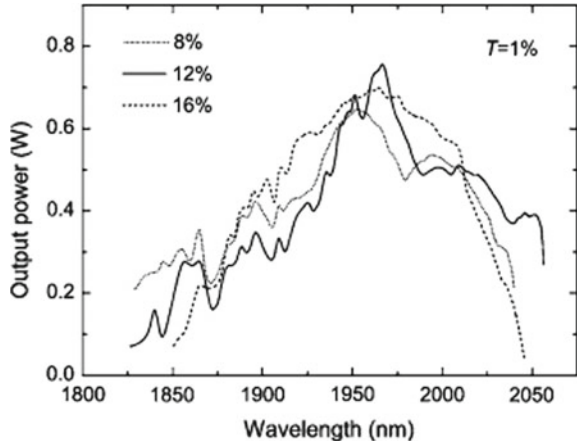


Fig. 5.14 Scheme of the Tm:LiLuF laser cavity and pump system [25]

Fig. 5.15 Tunability curves of Tm:LiLuF laser measured with 1% output coupling for Tm³⁺-doping level of 8, 12, and 16% [25]



The works we mention above represent high power results or significant large spectral tunability. More works with different hosts have been done using Lyot filter as a tuning component, and will be summarize later.

5.3.5 Febry-Perot Etalon Tuning

The Febry-Perot etalon wavelength selective transmission is used for narrowing and tuning laser emission. The etalon is composed from two reflective facets, usually with the same reflectivity. A plane wave incident on the etalon will have multiple reflection and transmission wave with relative phase between them. For specific wavenumber, the infinite series of reflected waves will have a destructive interference causing high transmission. For the ideal case of equal reflectivity, the maximum transmission is one.

The etalon transmission is given by the Airy formula:

$$T = \left[1 + \frac{4R \sin^2(\Gamma)}{(1 - R)^2} \right]^{-1} \tag{5.6}$$

where, T is the etalon transmission, R is the surface reflection and Γ is the single-pass phase shift in the etalon, which is given by:

$$\Gamma = \frac{2\pi}{\lambda} nd \cos(\theta) \tag{5.7}$$

Where λ is the laser wavelength, n is the refractive index of the etalon, d is the etalon thickness, and θ is the incident angle. Maximum transmission occurs when $\sin^2(\Gamma)$ is one, and is independent on the reflectivity degree. When compared to the phase

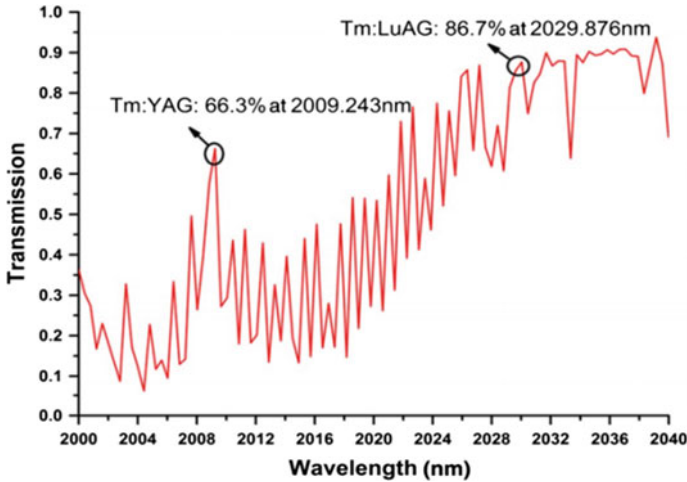


Fig. 5.16 Atmospheric transmission at 2–2.04 μm [27]

retardation of birefringent plate for normal incidence, the effective thickness of the etalon nd , is orders of magnitude thicker than the plate, $|n_o - n_e|d$. This implies that the free spectral range of even very thin etalon is much lower than birefringent plate. On the other hand, the wavelength selectivity of the etalon is much higher.

Most of the works of etalon inside laser cavity aimed for narrowing the linewidth of the laser down to single longitudinal mode. The tuning aspect of the etalon was a byproduct, and tunability range was small.

Wang et al. [27] published a single frequency Tm:LuAG using intra cavity etalons. Pair of two etalon with thicknesses of 1 and 0.1 mm was inserted to the cavity. The 0.1 mm etalon inserted to provide the free spectral range, and the 1 mm etalon was inserted to get linewidth narrowing. In order to achieve single frequency operation, the 1 mm etalon was coated for high reflectivity. By tuning the etalon angle, tunability of 11 nm was achieved. The Tm:LuAG tunability range locality enables better operation in the atmosphere due to higher transmission values, as can be seen in Fig. 5.16. The maximum single frequency output power was 616 mW at wavelength of 2021 nm, with a relative power stability of 0.61%. The wavelength tuning range of the single-frequency Tm:LuAG laser was 11 nm from 2018.714 to 2029.876 nm. The laser scheme and tunability results are shown in Fig. 5.17 and Table 5.1.

A work was done in our lab recently [28] used a 500 and 25 μm thicknesses YAG etalons. A high power Tm:YAP laser with linewidth smaller than 0.15 nm was achieved without coating on the etalons. Tunability range of 35 nm from 1917 to 1951 nm was achieved by tilting the etalon angle. The tunability range is smaller than what can be achieved using a Lyot filter, but it gave an added value of narrowing the linewidth and working with better polarization for every wavelength. Watt-level output power was measured along the laser tunable range, obtaining a maximal output power of 3.88 W at 1934 nm. A slope efficiency of 44.8% is demonstrated for

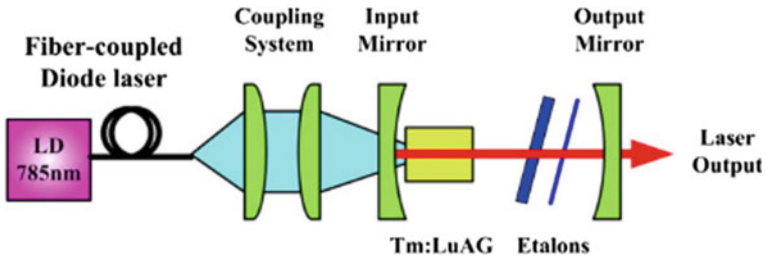


Fig. 5.17 Schematic diagram of single-frequency Tm:LuAG laser [27]

Table 5.1 Single-frequency output power at different wavelengths [27]

Wavelength (nm)	SLM output power (mW)	Wavelength (nm)	SLM output power (mW)
2018.714	235	2024.293	398
2018.694	309	2024.366	509
2018.845	545	2024.408	550
2020.127	360	2025.74	540
2020.203	480	2025.776	515
2020.218	564	2027.126	529
2021.452	308	2028.083	328
2021.498	348	2028.373	238
2021.583	616	2028.436	362
2021.627	598	2028.55	504
2022.988	588	2028.555	575
2023.006	605	2029.876	382

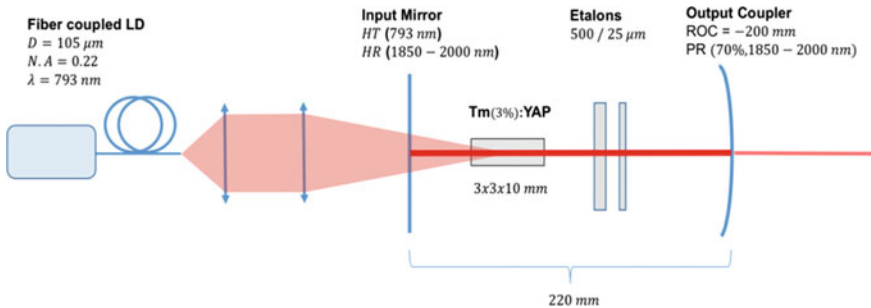
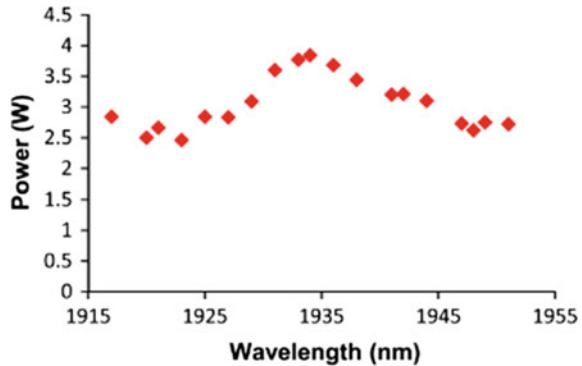


Fig. 5.18 Schematic of the Tm:YAP experimental setup [28]

a maximal absorbed pump power of 12.1 W. The laser Scheme is similar to Fig. 5.17, with different gain medium and pump wavelength, and is shown in Fig. 5.18, and the tunability results are shown in Fig. 5.19.

Fig. 5.19 Tunability curves of Tm:YAP laser using two etalons [28]



5.4 Pulsed Laser Tunability

Lasers can operate in a continuous or pulsed regime. In some cases, pulsed lasers are valuable when peak power rather than average power is most important. One of the most applied techniques to producing high energy pulses is the Q-switch technique. This technique allows the production of light pulses with extremely short duration and high peak power, much higher than can be produced by the same laser operating in continuous wave mode. The basic of Q-switch operation is varying the Q parameter of the laser cavity. High losses are inserted to the cavity preventing laser operation, and hence allowing the population inversion increase. If the losses are removed quickly, the extremely high stored population produces high energy short pulse.

There are two main categories of Q-switching—active and passive. For active Q switching the losses are modulated with an active control element. It is usually an acousto-optic (AOM) or electro-optic modulator (EO) where the pulse is formed shortly after an electrical trigger signal arrives. The achieved pulse energy and pulse duration depends on the gain medium properties, on the pump power and the pulse repetition rate. A passive Q-switch is using nonlinear behavior of an optical element, such as doped crystal, or a passive semiconductor device. The characteristic of such material is that transmission increases when the intensity of light exceeds some threshold (saturable absorbers). If such a material with high absorption at the laser wavelength is placed inside the laser resonator, it will initially prevent laser oscillation. As the gain increases during a pump pulse and exceeds the round-trip losses, the intra cavity power density increases dramatically causing the passive Q-switch to saturate. Under this condition the losses decrease fast and a Q-switch pulse builds up.

The pulse duration achieved with Q switching is typically in the nanosecond range, corresponding to several resonator round trips. The energy of the generated pulse is typically higher than the saturation energy of the gain medium and can be in the milli joule range even for small lasers. The peak power can be orders of magnitude higher than the power which is achievable in continuous-wave operation.

Q-switching has been demonstrated with many mid-IR solid-state materials with long upper-level lifetime.

Passive Q-switching techniques, based on saturable absorbers, require an appropriate absorbing material with low saturation intensity at the lasing wavelength. Such materials that were using in the 2 μm regime are: CrZnSe [29, 30], CrZnS [30, 31], Co^{2+} doped silver halide saturable absorber [32] or graphene [33].

Generally speaking, it could be argued that the active methods allow a better control on the pulse characteristics, while the passive methods are simple, compact and inexpensive.

In the case of tuning laser, the pulsed regime is relatively harder. The high gain in the pulse formation is hard to control with the small losses introduce by most of the techniques mention above. For passive Q-switch, the tuning should be easier by theory, because high gain starts after laser radiation is occurs, in contrast to the active methods. For the passive method, the laser threshold increases compared to the CW operation, especially for high initially loss of the SA. When tuning the laser, the selected wavelength's emission cross section becomes lower and also contributes to the threshold increasing. This combination can decrease the tuning range below the CW operation significantly when the pump power is limited.

There are few evidences of unsuccessful attempts to get continuously tuning in pulsed lasers. Jabczynski have reported [34] a CW tunability with Lyot filter but was unsuccessful to tune when the laser was Q-switched. Cole and Goldberg [31] have reported high energy pulsed laser with Lyot filter, but the tuning was between distinct peaks of the gain and not continues. Few other works implement a tunable pulsed laser by restricting themselves to low energy pulses.

A tunable Tm:LSO laser actively Q-switched using acousto-optic modulator (AOM) was published by Feng et al. [35]. The laser was tuned by a birefringent plate, and has tuning range of 111 nm. The pulse duration was few hundreds ns and maximum pulse energy of 0.17 mJ with the plate inside. The X shape cavity permit small beam spot on the crystal and on the AOM, reducing the switching time. Also the astigmatism produced by the non-normal incident can be compensated. The laser scheme can be shown in Fig. 5.20, and the laser tunability results are shown in Fig. 5.21.

Another work done with Tm:YAP using graphene as a passive Q-switch SA has achieved tunability [36]. The graphene is unique semiconductor used as a SA with very little wavelength dependent, in contrast to doped crystal, and with low non-saturable losses. Its saturation energy is also low, causing low energy and high repetition rate pulses. The tuning method was using a VBG as a folding mirror, and adjusting the VBG and the end mirror angles. The maximum pulse energy was 10.1 μJ and the tuning range was 35.5 nm. The laser Scheme is shown in Fig. 5.22, and the tunability results are shown in Fig. 5.23.

Recently a high energy (mJ level) tunable active Q-switch (AOM) Tm:YLF was built in our lab [37]. The tuning was done using an etalon pair providing higher selectivity than other tuning methods with a narrow bandwidth of 0.15 nm. The tuning range of 33 nm between 1873 and 1906 nm, and maximum pulse energy of 1.97 mJ was achieved at 1 kHz repetition rate. The use of etalons does not restrict

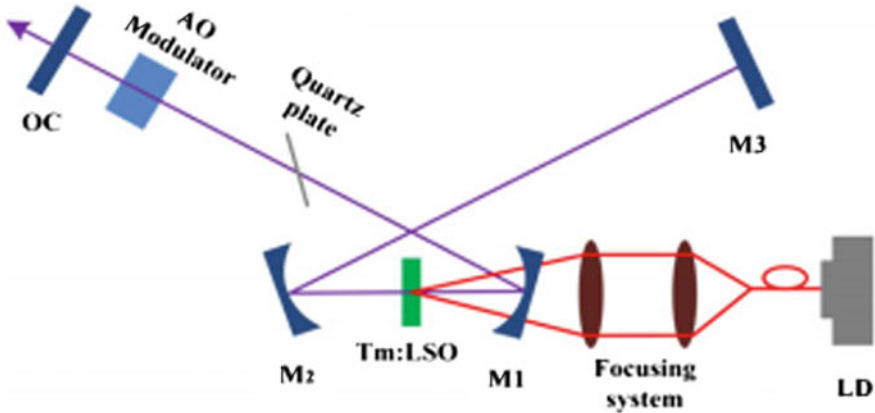


Fig. 5.20 Experimental setup of the actively AOM Q-switched Tm:LSO laser [35]

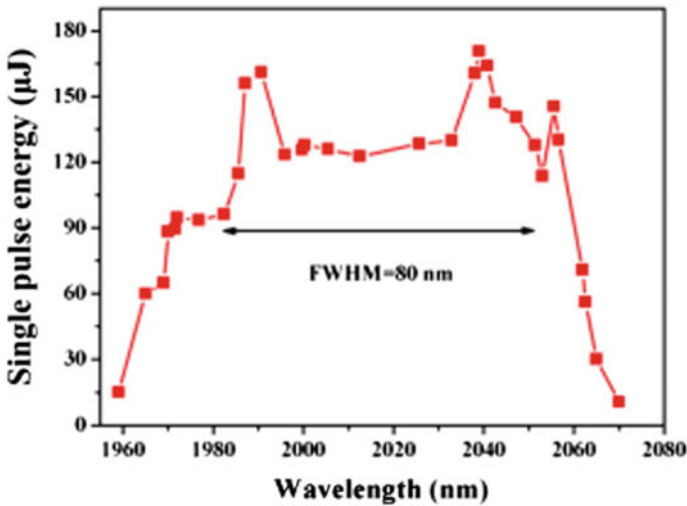


Fig. 5.21 Single-pulse energy versus output wavelength under an absorbed pump power of 3.85 W [35]

the laser cavity design or the polarization. It can be seen that both the 1906 and 1880 nm peaks are shown clearly, in contrast to tuning using Lyot filter (compare to the tunability results in Fig. 5.10). The linear resonator laser Scheme is shown in Fig. 5.24, and the tunability results are shown in Fig. 5.25.

As far as we know, this is the only Tm based laser work that achieved mJ level output energies over significant range of tunability. We assume that the high selectivity of the etalon has allow us to achieve that result.

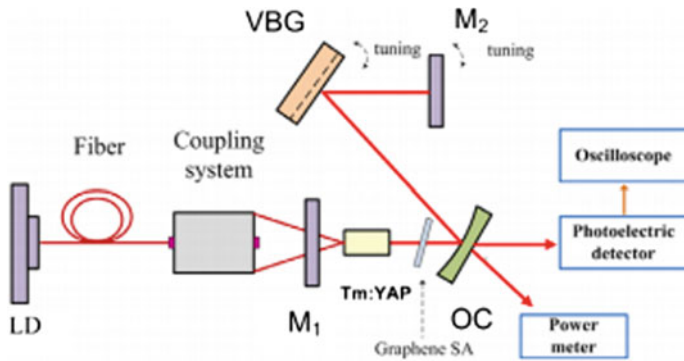


Fig. 5.22 Schematic diagram of the experimental setup [36]

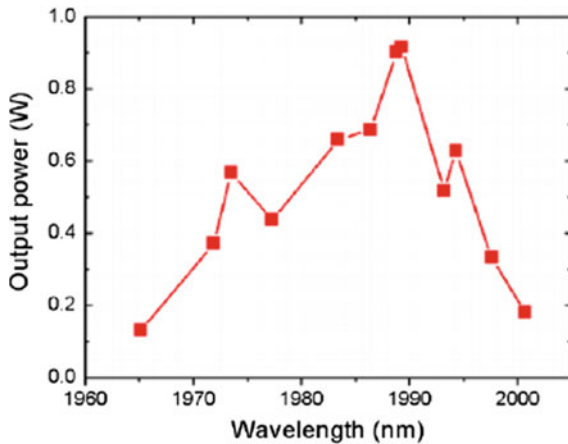


Fig. 5.23 Wavelength tunability of Q-switched Tm:YAP laser at the pump power of 12.6 W [36]

5.5 Tunability Tm Lasers Overview

A short summation of the works done on tunable Tm lasers is given in Table 5.2. Works done before 2007 can be found at Godard [2]. In this review we focused on works done since then. As can be seen, many works have been done demonstrating new hosts for Tm laser. This reflects the growing interest in 2 μm spectral regime especially in Tm lasers.

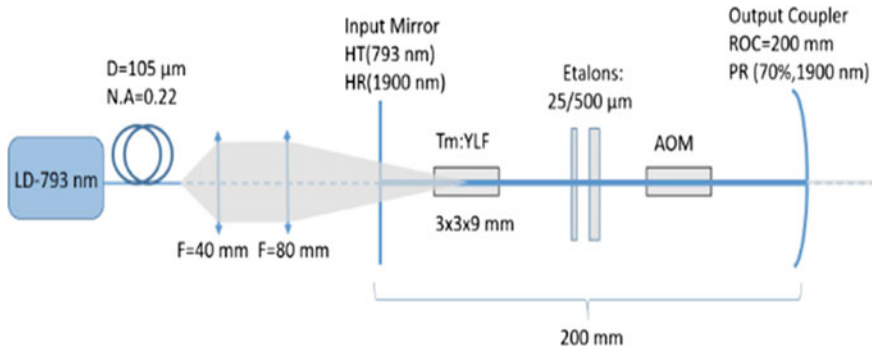


Fig. 5.24 Scheme of the Tm:YLF pulsed experimental setup [37]

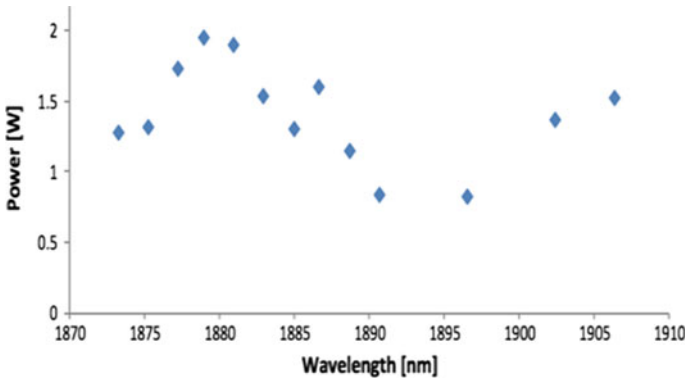


Fig. 5.25 Tunability curves of Tm:YLF laser using etalons [37]

Table 5.2 Tunability in Tm-doped lasers

Author	Year	Active crystal	Range (nm)	Maximum power	Tunability method
Qin et al. [11]	2013	Tm:CYA	1861–2046	90 mW	Prism
Gao et al. [13]	2013	Tm:CLNGG	1896–2069	181 mW	Prism
Lancaster et al. [14]	2014	Tm:ZBLAN	1745–1998	60 mW	Grating
Yu-Feng et al. [38]	2007	Tm:YAP	1899–2025	225 mW	BF
Yu-Feng et al. [25]	2007	Tm:GdVO ₄	1820–1946	2.5 W	BF
Coluccelli et al. [26]	2007	Tm:LiLuF	1817–2056	700 mw	BF
Němec et al. [34]	2009	Tm:YLF	1845–1935	0.6 W	BF

(continued)

Table 5.2 (continued)

Author	Year	Active crystal	Range (nm)	Maximum power	Tunability method
Eremeikin et al. [21]	2010	Tm:YLF	1860–1940	2.6 W	BF
Feng et al. [39]	2013	Tm:LSO	1936–2081	201 mW	BF
Trapani et al. [40]	2014	Tm:LuVO ₄	1843–1981	730 mW	BF
Feng et al. [35]	2014	Tm:LSO	1959–2070	170 mW	BF ^a
Feng et al. [41]	2014	Tm:SSO	1922–2020	340 mW	BF
Sun et al. [18]	2012	Tm:LuYAG	1935–1995	600 mW	VBG
Jin et al. [36]	2015	Tm:YAP	1965–2001	920 mW	VBG ^a
Zhao et al. [19]	2017	Tm:CALGO	1919–1994	5.8 W	VBG
Zhang et al. [42]	2009	Tm:YAG	2011–2016	514 mW	Etalon
Wang et al. [27]	2013	Tm:LuAG	2018–2029	616 mW	Etalon
Sheintop et al. [28]	2018	Tm:YAP	1917–1951	3.84 W	Etalon
Sheintop et al. [37]	2018	Tm:YLF	1873–1906	1.97 W	Etalon ^a

^aPulsed lasers

5.6 Conclusions

This work reviews the tunability aspect of solid state Tm based lasers. The characteristic wide fluorescence range around 2 μm of these lasers allows extensive tunability in CW mode and in Pulse mode. We introduced the main and different Tm hosts and there influence on the laser performance. Different tuning methods were theoretically described, and their advantages and drawbacks were presented. each method was introduce with remarkable results in order to give to the reader an update picture of what have been done in this field. From the relatively mature and widely implemented CW Tm lasers, we come to discuss the developing area of pulsed tunability. We showed the limitation of the works done and shown the possibility of achieving significant tuning combined with high energy pulses by using etalons, although it is less common tuning method. We believe that tuning high energy pulses at the 2 μm regime can open the way to improve many applications in various areas and have yet to reach its full potential.

References

1. I. Sorokina, K. Vodopyanov, *Solid-State Mid-infrared Laser Sources* (Springer, Berlin, 2003), pp. 255–351
2. A. Godard, Infrared (2–12 μm) solid-state laser sources: a review. *Comptes Rendus Phys.* **8**(10), 1100–1128 (2007)
3. K. Scholle, P. Fuhrberg, P. Koopmann, S. Lamrini, *2 μm Laser Sources and Their Possible Applications* (INTECH Open Access Publisher, 2010), pp. 471–500
4. W. Koehner, *Solid-State Laser Engineering* (Springer, New York, NY, 2006)

5. O. Svelto, *Principles of Lasers* (Springer, New York, 2016)
6. S. So, J. Mackenzie, D. Shepherd, W. Clarkson, J. Betterton, E. Gorton, A power-scaling strategy for longitudinally diode-pumped Tm:YLF lasers. *Appl. Phys. B* **84**(3), 389–393 (2006)
7. F. Cornacchia, A. Toncelli, M. Tonelli, 2- μm lasers with fluoride crystals: research and development. *Prog. Quantum Electron.* **33**(2–4), 61–109 (2009)
8. E.F. Kustov, V.P. Petrov, D.S. Petrova, J.P. Udalov, Absorption and luminescence spectra of Nd^{3+} and Er^{3+} ions in monocrystals of CaYAlO_4 . *Phys. Status Solidi a* **41**, 379–383 (1977)
9. K. Naito, A. Yokotani, T. Sasaki, T. Okuyama, M. Yamanaka, M. Nakatsuka, S. Nakai, T. Fukuda, M. Timoshechkin, Efficient laser-diode-pumped neodymium-doped calcium–niobium–gallium–garnet laser. *Appl. Opt.* **32**(36), 7387 (1993)
10. F.J. Duarte, *Tunable Lasers Handbook* (Academic Press, San Diego, 1995), pp. 289, 291
11. Z.P. Qin, J.G. Liu, G.Q. Xie, J. Ma, W.L. Gao, L.J. Qian, P. Yuan, X.D. Xu, J. Xu, D.H. Zhou, Spectroscopic characteristics and laser performance of Tm:CaYAlO₄ crystal. *Laser Phys.* **23**, 105806 (4 pp.) (2013)
12. R. Moncorgé, N. Garnier, P. Kerbrat, C. Wyon, C. Borel, Spectroscopic investigation and two-micron laser performance of Tm^{3+} :CaYAlO₄ single crystals. *Opt. Commun.* **141**(1–2), 29–34 (1997)
13. W. Gao, G. Xie, J. Ma, M. Liu, P. Yuan, L. Qian, H. Yu, H. Zhang, J. Wang, J. Zhang, Spectroscopic characteristics and efficient laser operation of Tm:CLNGG disordered crystal. *Laser Phys. Lett.* **10**(5), 055809 (2013)
14. D. Lancaster, S. Gross, M. Withford, T. Monro, Widely tunable short-infrared thulium and holmium doped fluorozirconate waveguide chip lasers. *Opt. Express* **22**(21), 25286 (2014)
15. O. Efimov, L. Glebov, L. Glebova, K. Richardson, V. Smirnov, High-efficiency Bragg gratings in photothermorefractive glass. *Appl. Opt.* **38**(4), 619–627 (1999)
16. L.B. Glebov, High brightness laser design based on volume Bragg gratings. *Proc. SPIE* **6216**, 621601 (2006)
17. T. Chung, A. Rapaport, V. Smirnov, L. Glebov, M. Richardson, M. Bass, Solid-state laser spectral narrowing using a volumetric photothermal refractive Bragg grating cavity mirror. *Opt. Lett.* **31**(2), 229 (2006)
18. M. Sun, J. Long, X. Li, Y. Liu, H. Ma, Y. An, X. Hu, Y. Wang, C. Li, D. Shen, Widely tunable Tm:LuYAG laser with a volume Bragg grating. *Laser Phys. Lett.* **9**(8), 553–556 (2012)
19. Y. Zhao, F. Wu, X. Xu, D. Shen, High-power, widely tunable Tm:CaGdAlO₄ laser resonantly pumped by a Raman fiber laser at $\sim 1.7 \mu\text{m}$. *Opt. Eng.* **57**(02), 1 (2017)
20. B. Lyot, Un monochromateur a grand champ utilisant les interferences enlumière polarisée. *Compt. Rend.* **197**, 1593 (1933)
21. J.W. Evans, The birefringent filter. *J. Opt. Soc. Am.* **39**, 229–242 (1949)
22. A.L. Bloom, Modes of a laser resonator containing tilted birefringent plates. *J. Opt. Soc. Am.* **64**, 447–452 (1974)
23. S. Zhu, Birefringent filter with tilted optic axis for tuning dye lasers: theory and design. *Appl. Opt.* **29**(3), 410 (1990)
24. O. Ereimeikin, A. Savikin, K. Pavlenko, V. Sharkov, Diode-pumped tunable Tm:YLF laser for mid-infrared gas spectroscopy. *Quantum Electron.* **40**(6), 471–474 (2010)
25. L. Yu-Feng, Y. Bao-Quan, L. Yan-Mei, W. Yue-Zhu, J. You-Lun, Widely tunable cw diode-pumped 1.9- μm Tm:GdVO₄ laser at room temperature. *Chin. Phys. Lett.* **24**(3), 724–726 (2007)
26. N. Coluccelli, G. Galzerano, P. Laporta, F. Cornacchia, D. Parisi, M. Tonelli, Tm-doped LiLuF₄ crystal for efficient laser action in the wavelength range from 182 to 206 μm . *Opt. Lett.* **32**(14), 2040 (2007)
27. L. Wang, C. Gao, M. Gao, L. Liu, F. Yue, Diode-pumped 2 μm tunable single-frequency Tm:LuAG laser with intracavity etalons. *Appl. Opt.* **52**(6), 1272 (2013)
28. U. Sheintop, E. Perez, S. Noach, Narrow bandwidth tunable watt level Tm:YAP laser using two etalons, in *Proceedings of the 6th International Conference on Photonics, Optics and Laser Technology* (2018)

29. A. Korenfeld, D. Sebbag, U. Ben-Ami, E. Shalom, G. Marcus, S. Noach, High pulse energy passive Q-switching of a diode-pumped Tm:YLF laser by Cr:ZnSe. *Laser Phys. Lett.* **12**(4), 045804 (2015)
30. D. Sebbag, A. Korenfeld, U. Ben-Ami, D. Elooz, E. Shalom, S. Noach, Diode end-pumped passively Q-switched Tm:YAP laser with 185-mJ pulse energy. *Opt. Lett.* **40**(7), 1250 (2015)
31. B. Cole, L. Goldberg, Highly efficient passively Q-switched Tm:YAP laser using a Cr:ZnS saturable absorber. *Opt. Lett.* **42**(12), 2259 (2017)
32. H. Hecht, Z. Burshtein, A. Katzir, S. Noach, M. Sokol, E. Frumker, E. Galun, A. Ishaaya, Passive Q-switching of a Tm:YLF laser with a Co²⁺ doped silver halide saturable absorber. *Opt. Mater.* **64**, 64–69 (2017)
33. P. Gao, H. Huang, X. Wang, H. Liu, J. Huang, W. Weng, S. Dai, J. Li, W. Lin, Passively Q-switched solid-state Tm:YAG laser using topological insulator Bi₂Te₃ as a saturable absorber. *Appl. Opt.* **57**(9), 2020 (2018)
34. J. Jabczynski, Ł. Gorajek, W. Zendzian, J. Kwiatkowski, H. Jelínková, J. Šulc, M. Němec, High repetition rate, high peak power, diode pumped Tm:YLF laser. *Laser Phys. Lett.* **6**(2), 109–112 (2009)
35. T. Feng, K. Yang, S. Zhao, J. Zhao, W. Qiao, T. Li, L. Zheng, J. Xu, Broadly wavelength tunable acousto-optically Q-switched Tm:Lu₂SiO₅ laser. *Appl. Opt.* **53**(27), 6119 (2014)
36. C. Jin, D. Li, Y. Bai, Z. Ren, J. Bai, Wideband tunable graphene-based passively Q-switched Tm:YAP laser. *Laser Phys.* **25**(4), 045802 (2015)
37. U. Sheintop, E. Perez, S. Noach, Actively Q switched tunable narrow bandwidth milli-Joule level Tm:YLF laser. *Opt. Express*
38. L. Yu-Feng, J. You-Lun, Y. Bao-Quan, W. Yue-Zhu, S. Ubizskii, A laser-diode-pumped widely tunable single-longitude-mode Tm:YAP laser at room temperature. *Chin. Phys. Lett.* **24**(9), 2594–2596 (2007)
39. T. Feng, S. Zhao, K. Yang, G. Li, D. Li, J. Zhao, W. Qiao, J. Hou, Y. Yang, J. He, L. Zheng, Q. Wang, X. Xu, L. Su, J. Xu, Diode-pumped continuous wave tunable and graphene Q-switched Tm:LSO lasers. *Opt. Express* **21**(21), 24665 (2013)
40. F. Di Trapani, X. Mateos, V. Petrov, A. Agnesi, U. Griebner, H. Zhang, J. Wang, H. Yu, Continuous-wave laser performance of Tm:LuVO₄ under Ti:sapphire laser pumping. *Laser Phys.* **24**(3), 035806 (2014)
41. T. Feng, S. Zhao, K. Yang, G. Li, D. Li, J. Zhao, W. Qiao, L. Zheng, J. Xu, Q. Wang, X. Xu, L. Su, Study on characteristics of diode-pumped continuous-wave tunable and passively Q-switched Tm:SSO laser. *Appl. Phys. B* **117**(1), 177–182 (2014)
42. Y. Zhang, C. Gao, M. Gao, Z. Lin, R. Wang, A diode pumped tunable single-frequency Tm:YAG laser using twisted-mode technique. *Laser Phys. Lett.* **7**(1), 17–20 (2010)

Chapter 6

A Dual-Wavelength Widely Tunable C-Band SOA-Based Fiber Laser for Continuous Wave Terahertz Generation



Muhammad Umyy, Abdullah Hossain, Simeon Bikorimana
and Roger Dorsinville

Abstract There have been several breakthroughs in terahertz science and technology. Major recent advances have brought novel avenues for terahertz generation to the limelight. Research work pertaining to terahertz generation and its applications has expanded at an unprecedented rate. This encompassing review provides a detailed description of current means of terahertz generation with a special focus on optical sources. Finally, a specific recent development utilizing a novel dual-wavelength, C-band, SOA-based, fiber hybrid compound-ring resonator source is discussed. Boasting both single and dual-port operation, this source is utilized to excite a terahertz photo-mixer and achieves a difference wavelength tuning range of 20.42 nm (i.e., 2.51 THz) which corresponds to a terahertz radiation range of 0.8–2.51 THz at room temperature.

6.1 Introduction

The pursuit of efficient generation of terahertz (THz) has seen increased interest over the last two decades. The amount of publications and research works conducted

M. Umyy (✉)
New York City College of Technology, Brooklyn, USA
e-mail: maummy@citytech.cuny.edu

A. Hossain · S. Bikorimana · R. Dorsinville
The City College of New York, New York, USA
e-mail: ahossai12@citymail.cuny.edu

S. Bikorimana
e-mail: sbikori00@citymail.cuny.edu

R. Dorsinville
e-mail: rdorsinville@ccny.cuny.edu

© Springer Nature Switzerland AG 2019
P. Ribeiro and M. Raposo (eds.), *Optics, Photonics and Laser Technology 2018*, Springer Series in Optical Sciences 223,
https://doi.org/10.1007/978-3-030-30113-2_6

in this science has grown exponentially. There is a significant lack of feasible and efficient means of generating radiation in that region of the spectrum, especially for commercial applications. Much of the interest and advancements in THz science and technologies have stemmed from the natural crossover between electronic and optical perspectives. Systems operating in that area are only useful with guided waves which are typical of optical lasers.

There are numerous THz applications as proposed by earlier works such as increasing fog and haze visibility with reduced scattering (Rayleigh) losses, detecting internal defects of objects masked by opaque objects, high bandwidth communications, security, and medical imaging purposes as well. The vital principle exploited in such applications is the transparent nature of common materials or substances when incident to THz radiation [1]. The medical applications are numerous as well because the THz band has very low photon energy levels at the two extreme ends of the spectrum, which makes it safer relative to X-ray imaging technologies [2]. THz waves are also more suited for high-resolution sensing and imaging than their X-ray counterparts. Its non-ionizing nature makes it very attractive for exploitation in biomedical [3], security screenings [4], and quality control of food, drugs, and semiconductors [5, 6].

6.2 Various Sources for Terahertz Generation

There are numerous sources of THz radiation, optical and otherwise. There are thermal sources, electronic, under which falls solid-state and vacuum sources, optical sources, under which falls lasers themselves and other sources pumped by lasers (optical pumping via both continuous and pulsed lasers), and finally mechanical excitation. This work concerns mainly the optical sources of THz generation. However, briefly glossing over the numerous other sources before going into optical sources in further detail is appropriate given the review nature of this work.

Thermal sources of THz emission include both extraterrestrial (solar radiation) and otherwise [7]. Some of the sources used in laboratory setups are heated rods of SiC, also known as globar [8], and mercury lamps [9]. Electronic sources, specifically vacuum electronic sources, which are characteristically very high-power sources, include free-electron lasers (FEL) [10], backward wave oscillators [11], synchrotrons [12], gyrotrons [13], and travelling wave tubes [14]. Solid-state electronics include transistors (high frequency) [15] and diodes (Gunn diodes) [16]. Frequency multiplication is typically used to shift the electronic emissions which are initially in the sub-terahertz region into the terahertz range [17]. There is also mechanical excitation via unpeeling of adhesive tape [18] and the fracturing of sugar molecules [19]. The latter involves the transfer of charges between two newly created uneven surfaces.

The remaining part of this chapter will discuss in detail the optical means of THz generation and detection. Various optical sources are listed in the following section before delving further into each. Towards the end of this work, a novel optical source to generate and detect THz radiation is discussed.

6.3 Optical Sources for Terahertz Generation

There generally are two main avenues of optical generation which have subcategories. One method is to generate THz emission directly using certain types of lasers, while the second is to use optical pumping which entails an optical source which pumps a THz source. This pumped source generates THz radiation which is usually focused and collimated by reflective mirrors onto THz calibrated detectors.

As for the method of using lasers directly to generate THz radiation, there are various types of lasers that have been used. Some have been built from common semiconductors such as germanium [20] and silicon [21]. Such sources involve electrically pumped crystal lasers that possess low angular divergences [22]. Quantum cascade lasers have also been used though recently the main challenge hampering their widespread use is the temperature restrictions when operating such lasers [23]. Gas lasers, such as CO₂ lasers [24, 25] or Raman gas lasers have also been used before they were generally displaced by solid-state lasers [26]. In short, semiconductor, gas, solid state, and quantum cascade lasers are utilized.

Visible and near-infrared lasers (NIR) are used for optical pumping of THz sources. Both continuous-wave and pulsed laser systems can be used, and there are techniques that fall under each type of laser system for generating THz emission. Two continuous wave lasers may be mixed together, and their difference frequency results in emission within the THz range. Such a continuous wave beam (as well as pulsed laser beams) can also be used to pump photo mixers as well. Such lasers include diode lasers [27], dual-mode lasers [28], multi-mode lasers, and VECSEL lasers [29]. Employing pumped THz sources that are periodically inverted crystals [30] and/or with tilted orientation increases overall efficiencies [31].

Pulsed lasers are also used for pumping of antennae [32] and photoconductive switches [33]. It is even possible to produce THz radiation by optical pumping of mere air by ionizing the air particles to produce a plasma channel [34]. Other substitutes for air can also be employed, such as noble gases [35]. THz parametric oscillation in crystals such LiNbO₃ can also be used [36].

6.3.1 Pulsed Lasers Systems

A pulsed THz system begins with a femtosecond (fs) laser which produces a train of optical pulses (typical pulse duration of 100 fs with 100 MHz repetition rate). The pulse train is split via a beam splitter; one half is used to generate the THz radiation while the second half is used to gate a THz detector. This technique is used to sample THz radiation fields at arbitrary sample points in time with sub-picosecond resolution. The assumption is that the short laser pulses are identical to each other. Therefore, by measuring over a sufficient number of sample points in time, one can measure the overall shape of a single THz pulse. The measurement, which is done

over many consecutive pulses are combined to form the general shape of a single pulse, thus giving the average shape of the THz pulses.

This sort of sampling method is based on the vital and precise synchronization between the THz field and the optical pulses gating the detector. For THz generation, the replicability of THz pulses is vital; if pulses within a pulse train are changing faster than the measurement time window, the overall pulse waveform will not be measured properly. The measurement time interval must be at the least, the product of the number of measured THz values and the time spacing between consecutive pulses of a pulse train. Signal averaging increases the data acquisition time even further.

This method requires delay manipulation of the optical probe pulse at the gate relative to the generated THz pulses. This can also be done by using dual femtosecond lasers [37]. The repetition rate of each laser is locked to the other with a frequency offset output. One laser is used to generate the THz pulse while the other is used to gate the detector. The delay of the THz pulse, relative to the gating of the detector, sweeps automatically at a rate determined by the frequency offset between the two laser sources. The need for dual laser sources and the resultant bulkiness of such systems prove to be disadvantageous in commercial applications.

6.3.2 Continuous Wave Optical Laser Sources

In parallel with the rapid progress made in pulsed time-domain THz generation techniques, similar strides have been made with continuous wave (CW) systems as well. In this case, the main challenge lies in the development of new or improvement of existing THz sources and detectors. CW systems can also make use of photo-mixing by difference frequency mixing. CW excitation of photoconductive antennae (PCs) entails pumping by two CW near-infrared laser beams with slightly mismatched wavelengths. The THz radiation emitted is the result of the photo-mixing technique, specifically when the beat signal of the two slightly different frequencies excite the PC antenna.

To output two slightly mismatched optical beams, two CW laser sources can be used (as was also the case for the pulsed method). One beam is fixed and centered at a specific wavelength while the other is tunable over an interval whose minimum bound is the wavelength of the fixed laser source. Tuning the second wavelength manipulates the difference wavelength (between the dual-wavelength signals) over a range of several nanometers which covers a range of frequencies in the terahertz gap [38]. Previous works have used distributed feedback (DFB) laser diodes [39], QCLs [40], and group III-V lasers [41] to excite PCs for THz generation; however, such systems, as just mentioned, require multiple laser sources and become cost-ineffective. In this light, a recent work which has demonstrated the ability of producing dual-CW beams with frequency offset will be discussed towards the end of this review [2].

Although the technique of dual-source CW beam combination has yielded advancements in CW-THz systems, such systems suffer from noise mixing of the

CW beams and frequency drift. One method to alleviate uncorrelated noise from dual-source CW mixing is to use external modulators [42]. In contrast, a single laser source with dual-wavelength and multimode eliminates the need of such systems [43–46].

For commercial uses, however, table top, compact sources that have operational stability under normal conditions (room temperature) which can produce highly coherent and powerful THz radiation are much more desirable. These however are generally not easy to achieve with pulsed laser sources making them less accommodating than CW systems. Agility, high scanning speed, high resolution, and simplicity of data processing [4, 47] are among other attractive attributes that make CW systems preferable over their pulsed counterparts.

6.4 Resonance Absorption Based THz Sources

6.4.1 *Photoconductive Emitters*

Photoconductive emitters (antennae) have been in use since the 1980s and have improved substantially since then. It was demonstrated that low THz output power and conversion (optical to THz) efficiency can be improved by filtering out the bias field resultant from the excited carriers. Furthermore, it was also observed that higher optical powers did not necessarily generate higher THz outputs but instead lead to optical saturation [48, 49]. Higher optical powers tend to cause optical saturation. An alternative to avoid optical saturation is to increase the electric bias field (increase in DC voltage bias of the antenna). To overcome slow photo response time of conventional semiconductor antennae (discussed next), low-temperature growth (LTG) technique can be employed, for example, for GaAs to force shorter electron lifetimes [50].

6.4.2 *Semiconductor Surface Emitters*

Photonic absorption causes electrons of a semiconductor to be excited from the lower valence bands to the higher conduction bands which generates a current as dictated by Maxwell's equations (change in current density is proportional to the electric field radiance). To enhance this natural phenomenon, metallization of the semiconductor antenna is recommended to allow the application of a large DC bias to the antenna. This in turn greatly enhances the generated photocurrent and allows efficient coupling of THz radiation into free space [1]. The THz pulses induce short bursts of photocurrents which if averaged over many pulses, can be easily measured (as a delay between the optical and THz pulses). The pulse durations are dependent on the lifetime of the semiconductor conduction carriers as the peaks fall within the

THz range for emitters with picosecond and sub-picosecond carrier recombination times.

Under femtosecond pumping of semiconductor surface emitters, self-evidently, the surfaces of bulk semiconductors generate THz radiation due to ultrafast charge transport and resonant optical rectification (resonant OR). The charge transport, which we mentioned is responsible for the THz radiation from the surface of semiconductor emitters, can be driven by the surface electric field or by a difference in the electrons and holes mobilities [51, 52]. The latter is known as the photo-Dember effect and is dominant in narrow-gap semiconductors such as InAs and InSb [53].

6.4.3 Gas Ionization

THz radiation can also be generated by ionizing gas (laser-plasma) as mentioned in the introduction, via high-power laser pulses. There are many processes which fall under this category such as ponderomotive electron acceleration [54], plasma-vacuum boundary crossing by a laser-plasma accelerated bunch [55] and plasma density dynamics [56]. Furthermore, it was demonstrated that higher THz intensities can be achieved via gas-plasma excitation by both fundamental and second harmonic pulses [57]. It was even confirmed that the four-wave mixing process, although cannot completely describe the more complicated physical details of THz generation and detection through the above process, still can be used as a valid approximation [58, 59] under lower optical intensities [60]. Several advantages of THz generation via gas ionization are: the observed linear increase of THz fields with increased optical pulse energies, THz radiation occurs over a wide spectral range facilitated by the lack of phonon resonance and finally, gas media do not have inherent reflection losses.

6.4.4 Intra-molecular Charge Generation Via Photoinduction

Macroscopic currents based on photoinduced charge transfer along aligned molecules also radiate THz fields [61]. There are several methods to force molecules to align: molecules injection into a crystal lattice [62], oriented deposition on surfaces and application of external voltage bias (as done with photoconductive emitters) [63]. Microscopic currents, produced by laser pulse excitation, flowing through aligned molecules sum up to a macroscopic current; this macroscopic current in turn radiates a THz field.

6.5 Non-resonant Nonlinear Interactions Based THz Sources

The non-resonance methods generate THz radiation because of non-linear interactions of optical waves through the non-resonant second order susceptibility of the medium. This phenomenon can be exploited by the photo-mixing of two distinct optical CW beams with shifted frequencies; this is known as difference frequency generation, and it results in THz parametric oscillation. Another method is the non-resonant optical rectification of laser pulses of a broadband laser source. Even though the medium's optical axes may not be periodically inverted (as done in schemes discussed later), it still outputs collimated narrow-band coherent THz emission which can be constructively harvested from the bulk of the source media given certain conditions.

6.5.1 *Difference Frequency Generation*

Parametric interaction of two waves with dissimilar frequencies in media with non-resonance second-order susceptibilities results in a third wave with a frequency related to the difference of the first two. The amplitude of this third resultant wave is obviously at a maximum when the two source waves are in phase with each other. THz waves can be generated in both the forward and backward directions with respect to the excited medium; the phase mismatch in the forward direction can be ideally eliminated but this is not possible in the backward direction [64]. There are three requirements for effective difference frequency THz generation: high second-order susceptibility, low THz absorption, and phase matching. Non-linear effects of two-photon absorption discourages the use of certain semiconductor crystals in high-power pumping conditions in the near-infrared and visible ranges. This causes substantial THz losses from the generation of free carriers and higher optical absorption by the semiconductor crystal [65].

Returning to the issue of phase matching, this requirement has a direct correlation to the length of the crystal used for THz generation. The optical coherence length limits the active length in which the optical excitation pump can operate within the crystal. This limitation stems any further increase of THz output and narrowing of the THz spectral line which would have been otherwise possible with extended crystal lengths.

To alleviate this hindrance, the use of crystal structures with periodically inverted axes was proposed to give rise to quasi-phase-matching (QPM) conditions [66, 67]. This is simply the periodic variation of the sign of the second-order optical susceptibility within the crystal. The most prominent crystals known for QPM are periodically poled ferroelectric crystals with one-dimensional gratings. GaAs and GaP semiconductors show great promise for use with QPM THz generation under different frequency generation [68].

6.6 Non-resonant Optical Rectification

In many cases, it is sufficient to have a quasi-monochromatic high-power beam with a properly broad spectral bandwidth to serve as an optical pumping source. This contrasts with difference frequency generation where two completely distinct optical beams, albeit, with close enough wavelengths, are used to pump a medium for excitation. The spectral bandwidth of sub-picosecond lasers is conducive for THz generation under the difference frequency interactions of each pair of its own spectral components. In other words, difference mixing between all possible pairs of spectral components of a single source's femtosecond pulses add coherently to produce short bursts of THz radiation. This is the fundamental principle which underlies all tabletop schemes that rely on femtosecond pulse pumping of optically transparent non-linear materials. Thus, difference frequency interaction is said to occur under optical rectification of the femtosecond pulses [65].

This scheme relies on the non-resonance second-order susceptibilities of non-centrosymmetric media, typically inorganic crystals such as LiNbO_3 , ZnSe , ZnTe , CdS , GaAs , GaSe , ZnGeP_2 , CdTe , LiTaO_3 , and others. It is of paramount importance that such a process takes place in spectral regions far removed from the resonances of the non-linear medium. If this condition is not satisfied, the medium will not be transparent to the optical pump beam and will have to be restricted to a very minimal size.

While we have already discussed femtosecond pulse excitation of PC and semiconductor emitters, such processes are resonant OR schemes. As such, they cannot be attributed to the parametric processes; the strong pump and THz absorption forces these interactions to take place in the thin surface of the medium. However, non-resonant optical rectification, simply known as optical rectification (OR), takes place in the actual bulk of the media. It is characterized by near-instantaneous THz response and preserves phase information between the optical and THz waves and outputs collimated THz radiation. OR is very well suited for high THz frequencies and powers [65].

In general, such schemes tend to have poor conversion efficiencies. Phase matching conditions and any other factors required for constructive interference of THz waves generated from different portions from the bulk of the crystal are typically hard to satisfy. Furthermore, the issue of minimizing THz attenuation within the non-linear sample poses a few challenges. Various techniques have been proposed to mitigate THz attenuation within the pumped crystal.

6.6.1 Cherenkov Crystal Geometry

In electro-optic crystals, such as those mentioned in the previous sub-section, collinear phase matching, specifically between the THz phase and optical group velocities, is not possible [69]. This, in turn, reduces the coherence lengths of the

crystals with the backward coherence length of the crystal being several times smaller than the forward coherence length. Phase matching can somewhat be obtained using a different approach, via a Cherenkov geometry.

The Cherenkov geometry entails generating THz emission at an angle to the direction of the laser pump; this angle is known as the Cherenkov angle. There is spatial phase matching along the pump axis which allows for longitudinally large crystals as needed. In the perpendicular direction however, there still exists a phase mismatch which forces the crystal to be transversely several times smaller than the transverse coherence length. Typically, the coherence length is on the order of the THz wavelength thus sharp focusing of the laser beam into a spot of that diameter or even smaller is required [70].

This geometry allows for the collection of THz waves from the lateral side of the crystal while lowering absorption losses [71]. The THz beams will travel the smallest possible path if collected from the upper lateral side of the Cherenkov cut crystal. Although these are remarkable positives for this scheme, crystals cut in such a geometry do not allow higher optical pump powers. As mentioned, the optical beam is sharply focused, and this prevents increasing the optical pump power. Increasing the pump power can exceed the damage threshold of the non-linear crystal. Consequently, the resultant THz powers tend to be lower.

6.6.2 Excitation via Tilted Optical Pump Beams

From the standpoint of generating the most powerful THz pulses, the OR scheme which employs a tilted optical pulse front is considered to be one of the finest approaches. In this OR scheme, a single optical pump beam travels through a THz source (for example: Mg-doped LiNbO₃ crystal) at a variable angle to the propagation direction of its wavefront; the THz beams are emitted in the normal axis to the pump beam [72]. The question that arises is which angle is best for THz generation? Such an angle would be the angle in which the optical group velocity projected onto the THz beam direction equals to THz phase velocity which is as mentioned in the previous scheme, known as the Cherenkov angle.

This ensures that the necessary phase matching between each pair of optical spectral components and the resultant THz wave is satisfied. Moreover, the pulse tilt angle is, to some extent, dependent on the frequency of each spectral component [73]. This scheme does away with sharp focusing of the pump beam as required by the previous approach and this allows higher pump powers by increasing the transverse size of the pumped area. Because phase matching is satisfied in both the longitudinal and transverse axes, the THz waves generated from any part of the crystal, interfere constructively. This allows higher optical to THz wave conversion efficiencies for crystals with low THz absorption coefficients. It was shown that LiNbO₃ shows greatest potential for use with this scheme [74].

6.6.3 *Quasi-phase Matching*

THz waves can be generated under QPM conditions in periodically poled crystals; the walk-off effect between the optical and THz waves at distances larger than the coherence length are compensated for by the periodically alternating polarization axes in the crystal. The width of each THz pulse is inversely proportional to the length of the crystal. Periodically poled LiNbO₃ crystals (PPLN), allow high-power isolated narrow lines under OR of continuous-wave difference frequency optical beams [75, 76].

Higher efficiency but with the drawback of lower THz output is possible with QPM GaAs under OR of femtosecond pulses [77]. However, the non-linear absorption effects of such a structure dramatically increase its optical and THz absorption. Higher pump powers are not possible with this specific medium. Thus, PPLN still remains the most viable option due to their transparency with visible and near-infrared pumping [67] and advanced poling methods [78]. THz frequency tuning is possible by changing poling periods and crystal dispersion through temperature manipulation. At higher frequencies, however, QPM interaction becomes quite inefficient at room temperatures. This is one issue which makes such a technique unsuitable for widespread use.

High THz absorption in Lithium Niobate crystals can be addressed using non-collinear geometries where the optical wave vectors which are collinear under OR of femtosecond pulse excitation and the THz wave vector are at proper angles with each other. As explained before, this approach allows the THz waves to travel the shortest distance possible when being radiated from the lateral crystal surface but allows the full use of the longitudinal crystal length with many domains.

It is required that QPM in the longitudinal axis be pumped by a beam which is sharply focused in the transverse direction to the force a spot size smaller than the expected THz radiation wavelength. This can be bypassed however by using a tilted PPLN crystal which simply has its domains oriented at a certain angle with respect to its edges [70, 71]. Tilted PPLN crystals have been successfully used for surface emitted THz generation under difference frequency interaction of sub-nanosecond pulses.

6.7 **Novel Single Fiber Optical Laser Source Utilizing Difference Frequency Generation for Terahertz Emission**

As shown earlier, an essential component of many THz sources is a dual-wavelength laser pump. Typically, the generation of tunable dual-wavelengths at telecom optical wavelengths in a single laser source is achieved by using a photonic crystal fiber [79], FBG filters [80], a Fabry-Perot filter in conjunction with an optical band-pass filter [81], a fiber double-ring filter [82], and an array waveguide grating [83]. Most

of the aforementioned methods require the use of circulators. Moreover, AWGs are not continuously tunable (i.e., fixed wavelength channel separation) and their tuning range is limited to around 12 nm range. Thus, in the C-band, three different laser sources are used to achieve a wide tuning range of CW-THz radiation without any gaps [84].

Recently, a novel fiber hybrid compound-ring semiconductor optical amplifier (SOA) laser source utilizing passive dual-wavelength beam combination was demonstrated [85]. This resonator architecture consists of N branches each with an SOA. An N number of split equivalent beams circulate throughout the N branches and are amplified by N SOAs. This hybrid compound ring resonator structure absolves the system from the need of external amplifiers. This setup achieves continuous C-band tunability over a range of 20.42 nm at room temperature corresponding to 0.8–2.51 THz. The operational principle of this novel fiber laser source is the coherent beam combining method (via constructive interference) at the 3 dB fiber couplers based on the passive phase-locking mechanism of the low power SOA-based all-single-mode fiber hybrid compound-ring resonator [86].

6.7.1 Experimental Optical Source Architecture

The experimental setup is demonstrated in Fig. 6.1; the C-band SOA-based tunable fiber laser consists of two nested ring cavities (i.e., hybrid compound-ring resonator) with two broadband SLMs, one on either side of the compound ring resonator. The SLMs allow for both single and dual output port operation depending on the reflectivity settings which are adjusted via the polarization controllers in the SLM loop.

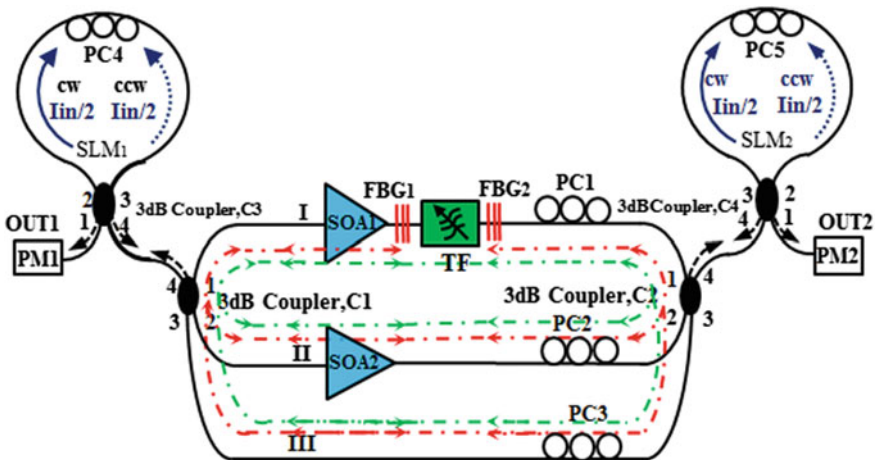


Fig. 6.1 Experimental setup of the dual-Sagnac loop mirror dual-wavelength SOA-based tunable fiber hybrid compound-ring laser [2]

Each ring cavity consists of two branches; Branches I and II for the inner ring cavity and Branches I and III for the outer ring cavity.

As evident, both cavities share Branch I which contains SOA₁ (Kamelian, OPA-20-N-C-SU), a tunable optical filter (TF-11-11-1520/1570) sandwiched between two similar FBGs centered at approximately 1551.97 nm, and a polarization controller, PC₁. Branch II contains SOA₂ (Thorlabs, S1013S), and polarization controller PC₂ while Branch III contains merely polarization controller PC₃ due to the lack of availability of further SOAs.

All branches are interconnected via two 3 dB fiber couplers, denoted C₁ and C₂, which also couple SLM₁ and SLM₂, correspondingly, to either side of the branches. SLM₁ and SLM₂ acts as a variable optical reflector whose reflectivity is manipulated by adjusting PC₄ and PC₅. Adjusting one SLM to maximum reflectivity and another to minimum reflectivity, allows for single port operation. Any other reflectivity setting in between the two extremes allows for dual-port operation.

Sandwiching the low power tunable optical filter which is tunable up to 1572.42 nm in the common branch, I, between two similar FBGs centered at 1551.98 nm allows for dual-wavelength selection. The three PCs (PC₁, PC₂, and PC₃) control the state of polarization of the light circulating within the compound ring cavity.

6.7.2 Dual-Wavelength Amplification and Single/Dual Output Port Operation

A tunable dual-wavelength beam was realized in the hybrid compound-ring resonator using two FBG filters, FBG₁ and FBG₂, of the same Bragg wavelength (i.e., $\lambda_{\text{FBG1}} = \lambda_{\text{FBG2}} = 1551.98$ nm) and a widely tunable optical filter from 1520 nm to 1570 nm. The fixed wavelength is determined by the FBGs, while the tunable wavelength, λ_{TF} , is obviously selected by the tunable filter denoted TF. The wavelength selection is performed in the common branch, I, of the hybrid compound-ring cavity.

Both SOAs are driven well above the threshold bias current level and the reflectivity of each SLM is adjusted to $\leq 0.1\%$. When the bias current of either SOA exceeds the total losses of the fiber compound-ring cavity, amplified spontaneous emission (ASE) is emitted from the SOAs and propagates in the forward and backward directions of the branches. The ASE from SOA₁ circulates in a clockwise direction within the ring cavity; it passes the FBG₁ filter, which reflects a fixed wavelength, λ_{FBG1} , back into SOA₁ for further amplification. The ASE signal continues to the tunable filter, TF, which filters out all wavelengths from except a single tunable wavelength λ_{TF} which is different from the wavelength fixed by the FBGs.

Subsequently, the beam with tunable wavelength, λ_{TF} , transverses through both the FBG₂ filter and polarization controller PC₁. After passing through PC₁, it is equally split and coupled by the 3-dB fiber coupler, C₂, where it is equally split. Specifically, the beam enters from Branch I via port 1 and is equally split into ports

2 and port 3 and thus coupled into Branches II and III of the fiber compound-ring cavity, respectively.

The beam circulating within Branch II passes through polarization controller PC_2 and is amplified by SOA_2 (i.e., specifically when the bias current level is above 180 mA). It then arrives at port 2 of the 3-dB fiber coupler C_1 , where it is coupled with the beam from port 3 which propagates through Branch III. Subsequently, this coupled signal is then equally split again between ports 1 and 4.

Half of the beam at port 4 of the 3-dB fiber coupler C_1 is fed into the output coupler, SLM_1 . Because the reflectivity of SLM_1 has been adjusted to $\leq 0.1\%$ as mentioned earlier, the selected beam with λ_{TF} exits at port 1 (i.e., OUT_1) of the 3-dB fiber coupler, C_3 . The other half is coupled back into Branch I via port 1 of the 3-dB fiber coupler C_1 and is further amplified by SOA_1 . This sequence of events completes and transverses the ring cavity in the clockwise direction which permits lasing of the tunable wavelength selected by the tunable filter λ_{TF} .

Now that the path that the tunable wavelength beam propagates through has been discussed, the fixed wavelength beam path must likewise be discussed. The fixed wavelength, denoted by λ_{FBG1} , is selected by FBG. It is reflected back to SOA_1 in the counter-clockwise direction where it is amplified. It is then equally split into two beams by the 3-dB fiber coupler C_3 and is coupled into Branches II and III of the fiber compound-ring cavity. The beam which propagates in Branch II is further amplified by SOA_2 but the beam in Branch III is not amplified as it lacks an SOA. Both beams propagate through polarization controllers, PC_2 and PC_3 respectively before being combined and equally split at ports 1 and 4 of the 3-dB fiber coupler C_2 .

The beam at port 4 of the 3-dB fiber coupler C_2 is fed into SLM_2 and then exits from port 1 (i.e., OUT_2) of 3-dB fiber coupler, C_4 . The other half of the beam with fixed wavelength λ_{FBG1} is coupled back into Branch I and travels to the second fiber Bragg grating filter, FBG_2 , and is reflected toward the 3-dB fiber coupler C_4 . Once again, it is equally split and respectively coupled into Branches II and III for further amplification which leads to lasing of the fixed Bragg wavelength, λ_{FBG} , after it completes a round trip back to the FBG_1 filter while undergoing further amplification through SOA_1 and SOA_2 .

Because this experimental setup does not utilize an optical circulator, both wavelengths of λ_{TF} and λ_{FBG1} circulate in both clockwise and counter-clockwise directions and finally emerge at the output couplers, SLM_2 and SLM_1 , respectively. Therefore, dual lasing wavelengths, the tunable λ_{TF} and fixed λ_{FBG} travel throughout the fiber hybrid compound-ring cavity and can be extracted at both output couplers, OUT_1 and OUT_2 . When output coupler SLM_1 is set to maximum reflectivity (i.e., $\geq 99.9\%$), this forces the dual wavelength beam to exit from the other output coupler SLM_2 . Similarly, if the output coupler SLM_2 is set to maximum reflectivity (i.e., $\geq 99.9\%$), this will force the dual wavelength beam to exit from the other output coupler SLM_1 . The wavelength separation (i.e., $\Delta\lambda_{THz}$) is adjusted by adjusting the tunable filter, TF. It is imperative to state that the path lengths of both loops are approximately equivalent to each other as all the branches of the hybrid ring cavity have identical length. Furthermore, all the fiber connections done in the entire optical system were done via FC/APC connectors.

6.7.3 Dual-Wavelength Optical Beam Tunability and Optical Power Stability

Initially, the bias currents for SOA₁ and SOA₂ were set to 200 and 500 mA, respectively. The reflectivity of SLM₁ was adjusted to minimum while the reflectivity of SLM₂ was set to maximum. This was done by adjusting the polarization controllers within the SLM loops. The dual-wavelength signal was measured with an OSA. The wavelength separation was manipulated by manual adjustment of the tunable filter. The tunable filter allowed the tunable wavelength to sweep from 1553.79 to 1572.34 nm.

While tuning the filter, the polarization controllers within the ring cavity branches, PC₁, PC₂, and PC₃, were optimized for each wavelength separation. The fixed wavelength was selected by two FBG filters which were centered at 1551.98 nm. Figure 6.2 shows the wavelength separation (i.e., beat signal: $\Delta\lambda_{\text{THz}} = \lambda_2 - \lambda_1$) of 3 nm and 20.35 nm corresponding to CW-THz beat frequencies (i.e., $\Delta\nu_{\text{THz}} = c * (\Delta\lambda_{\text{THz}}/(\lambda_1 * \lambda_2))$) of 0.37 and 2.5 THz, respectively.

The peak signals that were deduced from the measured output wavelength spectra, using an OSA (e.g., Fig. 6.2), were utilized to assess the optical signal-to-noise ratio

Fig. 6.2 Illustrates the wavelength spectrum of the dual-wavelength fiber hybrid compound-ring laser [2]

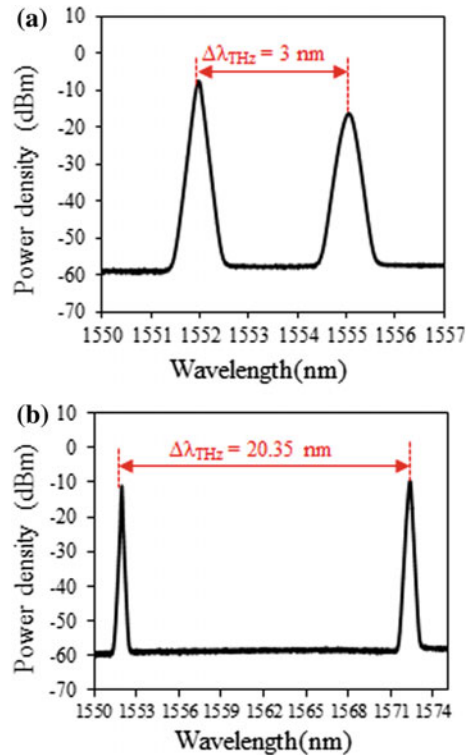
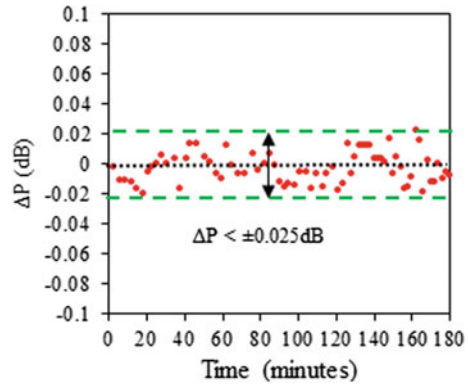


Fig. 6.3 Shows the output power short-term fluctuations measured over three hours at room temperature [2]



(OSNR). The peak power value at each center wavelength was subtracted from the background noise level of each wavelength spectrum. The OSNR for both the fixed wavelength and tunable wavelength remained significantly higher than +50 dB and +45 dB respectively over the entire tuning range.

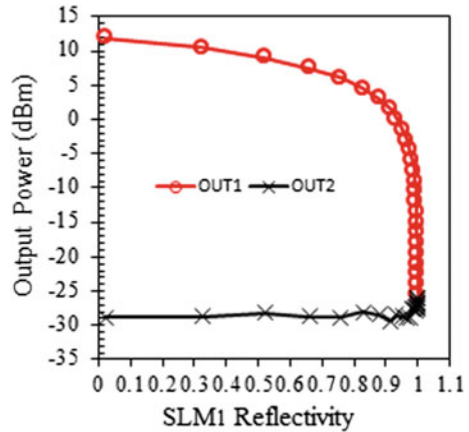
One of the important characteristics of any system is stability. Therefore, an optical power stability test at typical conditions (i.e., room temperature) was performed. SOA_1 and SOA_2 were set to the standard bias current levels of 200 and 500 mA, respectively. The optical power stability test lasted for a sizable duration of three hours with intervals of three minutes and an OSA resolution of 0.01 nm. Figure 6.3 records the power fluctuations of the proposed fiber hybrid compound-ring laser which were within ± 0.025 dB. Further reductions of observed fluctuations can be accomplished with proper and better packaging of the system.

6.7.4 Single and Dual Output Port Operation Characterization

The experimental tunable fiber laser can operate with two adjustable and switchable output ports, OUT_1 and OUT_2 . The output power can be adjusted from either port, as mentioned earlier, by manipulating the reflectivity of SLM_1 and SLM_2 but maintaining constant bias levels for SOA_1 and SOA_2 . In single port operation, only one SLM is set to maximum reflectivity, and the other is set to minimum reflectivity.

In order to characterize the power tunability of both ports, SLM_1 was set to minimum reflectivity while SLM_2 was set to maximize reflectivity. The initial output power measured from OUT_1 and OUT_2 was +11.85 dBm and -28.9 dBm respectively. By maintaining the reflectivity of SLM_2 to maximum and sweeping SLM_1 reflectivity from minimum to maximum by adjusting PC_4 , we obtained the characterization in Fig. 6.4. As evident, the output power from OUT_1 swept from +11.85 to -28.5 dBm as the SLM_1 reflectivity was changed. This demonstrates the single

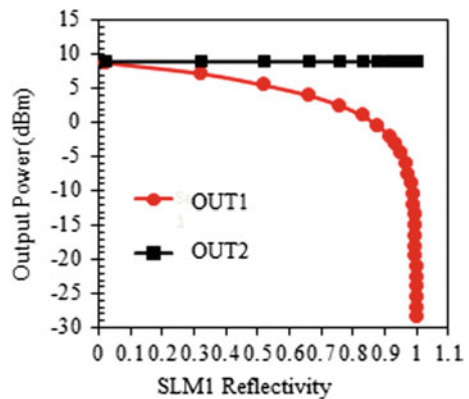
Fig. 6.4 Shows the output power from both output ports, OUT1 (unfilled circles) and OUT2 (crosses) as a function of different reflectivity values of the Sagnac loop mirrors for single output port operation [2]



output port operation. It should be noted that reversing the initial reflectivity settings for both SLMs is also possible for single output port operation.

Regarding the characterization of dual port operation, both ports were set to minimum reflectivity at first which corresponded to +8.94 dBm and +8.95 dBm for SLM₁ and SLM₂, respectively. By gradually increasing the reflectivity of SLM₁, and thus the power from OUT₁ from +8.94 to -28.9 dBm, while adjusting the SLM₂ reflectivity simultaneously to keep the OUT₂ optical power at +8.95 dBm, the curves in Fig. 6.5 were obtained. SLM₂ reflectivity was adjusted to about 50% to keep its output power from OUT₂ fixed at +8.95 dBm while SLM₁ was being adjusted from 0.1% to 99%, respectively.

Fig. 6.5 Illustrates the output power from both output ports, OUT1 (filled circles) and OUT2 (filled squares) for different reflectivity values of the Sagnac loop mirror, SLM1 for dual-output port operation while maintaining constant output power at OUT2 constant at +8.95 dBm via SLM₂ [2]



6.7.5 Continuous Wave THz Emission Generation and Detection

The experimental tunable fiber laser served as an optical source to excite an off-the-shelf photo-mixer (CW-THz based Indium Gallium Arsenide (InGaAs)) as an emitter and a pyroelectric-based terahertz sensor as a detector. The complete system setup that was utilized for the generation and detection of CW-THz radiation is shown in Fig. 6.6.

SOA₁ and SOA₂ was kept at constant bias levels of 250 and 500 mA, respectively. The optical power of the fiber laser source was maximized via adjustment of the polarization controller PC₅ in SLM₂. A wideband 1 × 2 single-mode fiber optic coupler with a split ratio of 95% and 5% was connected to the output coupler, OUT1, of the fiber laser system. The 95% branch of the optical splitter was fed to the CW-THz InGaAs emitter (TX) to serve as the excitation source. The 5% branch of the optical splitter was connected to an inline power meter (PM) to observe the optical power for the excitation source. An OSA was employed to observe the output spectrum of the dual-wavelength laser. The tunable filter was manually incremented by 0.1 nm. To further improve this resolution, however, an automated continuously tunable filter should be used. The optical beat signal from the fiber laser source was incident on the terahertz InGaAs-photomixer via a single-mode fiber and FC/APC connector.

The bulk of the terahertz radiation emitted from the antenna was radiated through an Indium Phosphide (InP) substrate. For proper coupling of this THz radiation into free space, a silicon lens was used. By using an off-axis parabolic mirror, OPM, the THz radiation from the emitter was collected and collimated. A polytetrafluoroethylene (i.e., Teflon) lens was then used to focus the THz signal onto the pyroelectric-based terahertz detector. The THz detector, as mentioned in the system specifications, was calibrated from 0.8 to 30 THz (Gentec-EO, Inc). An optical chopper provided a reference signal of 25.7 Hz chopping frequency to the THz detector module for increased signal-to-noise ratio.

In order to increase the THz radiation emission, the InGaAs-photomixer was biased at a DC voltage of −1.4 V. This photo mixer has a carrier lifetime of 0.3 ps and bandwidth of 3 THz [87]. The emitted THz radiation sweeps from 0.875 to 2.51 THz when the wavelength is tuned from 1554 to 1572 nm. The filter was adjusted in step intervals of 0.5 nm. This corresponds to a range of 1.2 GHz within the C-band. The maximum measured average power was around 350 nW, as shown in Fig. 6.7. The theoretical curve was drawn via (1) below [88]:

$$P_{\text{THz}}(\omega) \approx \frac{A}{1 + (\omega\tau)^2} \quad (1)$$

Fig. 6.6 **a** Pictorial view of the experimental setup for CW-terahertz radiation measurements (TX: InGaAs-photomixer and RX: pyroelectric THz sensor) and **b** Its schematic diagram [2]

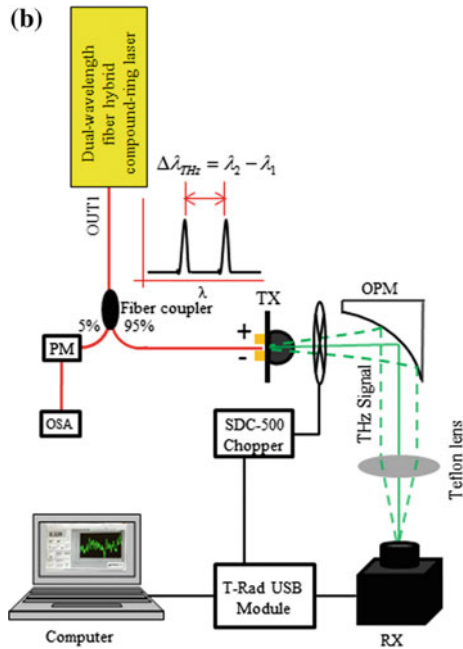
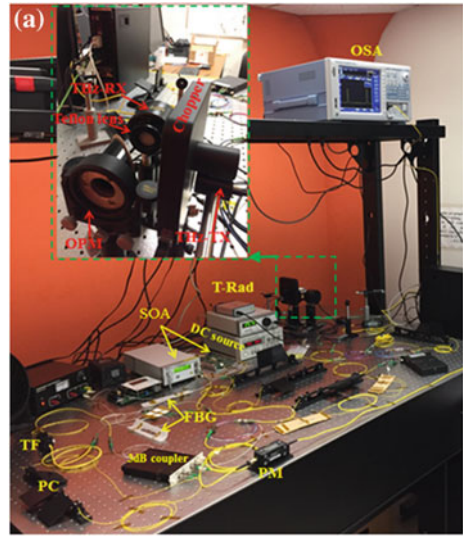
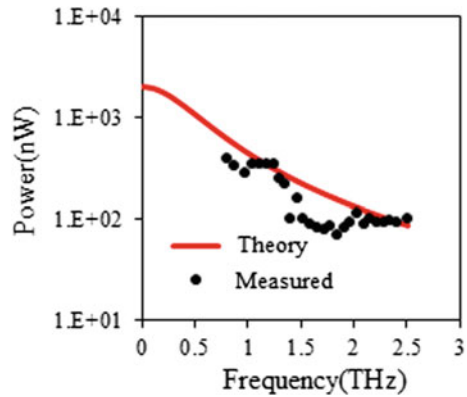


Fig. 6.7 Frequency spectrum of CW-terahertz radiation measured by a pyroelectric terahertz detector and theoretical curve fit $A = 2 \mu\text{W}$ and $\tau = 0.3 \text{ ps}$ [2]



6.8 Conclusion

In summary, a brief review of optical-based sources for THz radiation generation was presented. The field of optically generated THz radiation is an ever evolving and rapidly advancing field. As such, any possible conclusions tend to be short lived. There are numerous methods proposed for THz generation, each exploiting unique natural properties of different source media. Both new and older generation techniques can be brought to the forefront based on constantly developing ideas. There are many recently proposed techniques to minimize the footprint of optical systems being used for THz generation. As the main objective of such research works is to perfect THz generation methods for commercial applications, the focus of most research works currently is on higher efficiency methods and THz power output, and lower system footprints and costs.

The spectral range of CW methods are limited in all schemes due to the phonon absorption in non-resonance schemes and carrier response times for resonance methods. The advantages and disadvantages of resonance optical methods, which include charge transfer within photoconductive emitters, semiconductor surface emitters, and ionized gases, and non-resonance methods such as optical rectification, difference frequency generation, quasi-phase-matching, tilted pulse, Cherenkov geometry were discussed.

Furthermore, a novel approach to an optical source using low power optical components via a dual-wavelength C-band SOA-based fiber hybrid compound-ring laser used for the generation of continuous-wave terahertz radiation was discussed. The laser was used to excite a terahertz InGaAs photo mixer which then in turn emitted THz radiation which was subsequently detected by a pyroelectric terahertz sensor. The largest achieved tuning range, $\Delta\lambda_{\text{THz}}$, of 20.42 nm, corresponds to a CW-THz beat signal (i.e., $\Delta\nu_{\text{THz}}$) of 2.51 THz. CW tuning was achieved at room temperature from a single light source which is a primary advantage over systems that require multiple optical sources to achieve the same wavelength tuning range while minimizing gaps in the tuning range.

To further improve on the current system, an $N \times N$ hybrid compound-ring structure with N -number of SOAs can be used. Furthermore, by employing FBG filters which are centered much farther away towards the beginning of the C-band spectrum at approximately 1520 nm, one can achieve a wide range of 50 nm. This improved fiber laser system can excite an electro-optic crystal [79] to generate beat signals, $\Delta\nu_{\text{THz}}$, that can potentially reach up to 6 THz.

References

1. W. Chan, J. Deibel, D. Mittleman, Imaging with terahertz radiation. *Rep. Prog. Phys.* **70**, 1325–1379 (2007)
2. M.A. Ummy, S. Bikorimana, A. Hossain, R. Dorsinville, Continuous tunable terahertz wave generation via a novel CW optical beat laser source, in *Paper presented at the 6th International Conference on Photonics, Optics, and Laser Technology*, Funchal, Madeira, Portugal, 25–27 Jan 2018
3. E. Pickwell, V. Wallace, Biomedical applications of terahertz technology. *J. Phys. D Appl. Phys.* **39**(17), R301–R310 (2006)
4. N. Karpowicz, H. Zhong, J. Xu et al., Comparison between pulsed terahertz time-domain imaging and continuous wave terahertz imaging. *Semicond. Sci. Technol.* **20**(7), S293 (2005)
5. B. Hu, M.C. Nuss, Imaging with terahertz waves. *Opt. Lett.* **20**(16), 1716–1718 (1995)
6. M. Yamashita, K. Kawase, C. Otani, T. Kiwa, M. Tonouchi, Imaging of large-scale integrated circuits using laser terahertz emission microscopy. *Opt. Express* **13**(1), 115–120 (2005)
7. C. Thacker, A. Cooray, J. Smidt et al., H-ATLAS: the cosmic abundance of dust from the far-infrared background power spectrum. *Astrophys. J.* **768**(1), 58 (2013)
8. V.M. Zolotarev, R.K. Mamedov, A.N. Bekhterev, B.Z. Volchek, Spectral emissivity of a global lamp in the 2–50- μm region. *J. Opt. Technol.* **74**, 378–384 (2007)
9. K. Charrada, G. Zissis, M. Aubes, Two-temperature, two-dimensional fluid modelling of mercury plasma in high-pressure lamps. *J. Phys. D Appl. Phys.* **29**, 2432–2438 (1996)
10. B.A. Knyazev, G.N. Kulipanov, N.A. Vinokurov, Novosibirsk terahertz free electron laser: instrumentation development and experimental achievements. *Meas. Sci. Technol.* **21**, 054017 (2010)
11. M. Mineo, C. Paoloni, Corrugated rectangular waveguide tunable back-ward wave oscillator for terahertz applications. *IEEE Trans. Electron Devices* **57**, 1481–1484 (2010)
12. J.M. Byrd, W.P. Leemans, A. Loftsdottir, B. Marcellis, M.C. Martin, W.R. McKinney, F. San-nibale, T. Scarvie, C. Steier, Observation of broadband self-amplified spontaneous coherent terahertz synchrotron radiation in a storage ring. *Phys. Rev. Lett.* **89**, 224801 (2002)
13. V.L. Bratman, Y.K. Kalynov, V.N. Manuilov, Large-orbit gyrotron operation in the terahertz frequency range. *Phys. Rev. Lett.* **102**, 245101 (2009)
14. S. Bhattacharjee, J.H. Booske, C.L. Kory et al., Folded waveguide traveling-wave tube sources for terahertz radiation. *IEEE Trans. Plasma Sci.* **32**, 1002–1014 (2004)
15. J. Lusakowski, W. Knap, N. Dyakonova et al., Voltage tuneable terahertz emission from a ballistic nanometer InGaAs/InAlAs transistor. *J. Appl. Phys.* **97**, 064307 (2005)
16. S. Pérez, T. González, D. Pardo, J. Mateos, Terahertz Gunn-like oscillations in InGaAs/InAlAs planar diodes. *J. Appl. Phys.* **103**, 094516 (2008)
17. J.V. Siles, J. Grajal, Physics-based design and optimization of Schottky diode frequency multipliers for terahertz applications. *IEEE Trans. Microw. Theor. Technol.* **58**, 1933–1942 (2010)
18. E.N. Harvey, The luminescence of adhesive tape. *Science* **89**, 460–461 (1939)
19. E.N. Harvey, The luminescence of sugar wafers. *Science* **90**, 35–36 (1939)
20. D.S. Ong, H.L. Hartnagel, Generation of THz signals based on quasi-ballistic electron reflections in double-heterojunction structures. *Semicond. Sci. Technol.* **22**, 981 (2007)

21. M.A. Odnoblyudov, A.A. Prokofiev, I.N. Yassievich, K.A. Chao, Theory of a strained p-Ge resonant-state terahertz laser. *Phys. Rev. B* **70**, 115209 (2004)
22. Y. Chassagneux, R. Colombelli, W. Mauneult, S. Barbieri, H.E. Beere, D.A. Ritchie, S.P. Khanna, E.H. Linfield, A.G. Davies, Electrically pumped photonic-crystal terahertz lasers controlled by boundary conditions. *Nature* **457**, 174–178 (2009)
23. B.S. Williams, Terahertz quantum-cascade lasers. *Nature Photon.* **1**, 517–525 (2007)
24. F. Klappenberger, K.F. Renk, Transient-pulse nonlinear spectroscopy with the radiation of a multimode THz gas laser. *Int. J. Infrared Millimeter Waves* **24**, 1405 (2003)
25. V.A. Gorobets, B.F. Kuntsevich, V.O. Petukhov, Absorption by CS₂ molecules on “hot” band emission lines from a TEA CO₂ laser. *J. Appl. Spectrosc.* **74**, 67 (2007)
26. G. Dodel, On the history of far-infrared (FIR) gas lasers: Thirty-five years of research and application. *Infrared Phys. Technol.* **40**, 127–139 (1999)
27. K.A. McIntosh, E.R. Brown, K.B. Nichols, O.B. McMahan, W.F. DiNatale, T.M. Lyszczarz, Terahertz photomixing with diode lasers in low-temperature-grown GaAs. *Appl. Phys. Lett.* **67**, 3844 (1995)
28. M. Tani, P. Gu, M. Hyodo, K. Sakai, T. Hidaka, Generation of coherent terahertz radiation by photomixing of dual-mode lasers. *Opt. Quantum Electron.* **32**, 503–520 (2000)
29. M. Scheller, J.M. Yarborough, J.V. Moloney, M. Fallahi, M. Koch, S.W. Koch, Room temperature continuous wave milliwatt terahertz source. *Opt. Express* **18**, 27112–27117 (2010)
30. K.L. Vodopyanov, Optical generation of narrow-band terahertz packets in periodically inverted electro-optic crystals: conversion efficiency and optimal laser pulse format. *Opt. Expr.* **14**, 2263–2276 (2006)
31. J.A. Fülöp, L. Pálfalvi, G. Almási, J. Hebling, Design of high-energy terahertz sources based on optical rectification. *Opt. Express* **18**, 12311–12327 (2010)
32. D.H. Auston, Picosecond optoelectronic switching and gating in silicon. *Appl. Phys. Lett.* **26**, 101–103 (1975)
33. Y.C. Shen, P.C. Upadhyaya, E.H. Linfield, H.E. Beere, A.G. Davies, Ultrabroadband terahertz radiation from low-temperature-grown GaAs photoconductive emitters. *Appl. Phys. Lett.* **83**, 3117–3119 (2003)
34. M. Kress, T. Löffler, S. Eden, M. Thomson, H.G. Roskos, Terahertz-pulse generation by photoionization of air with laser pulses composed of both fundamental and second-harmonic waves. *Opt. Lett.* **29**, 1120–1122 (2004)
35. Y. Chen, M. Yamaguchi, M. Wang, X.C. Zhang, Terahertz pulse generation from noble gases. *Appl. Phys. Lett.* **91**, 251116 (2007)
36. J.I. Shikata, M. Sato, T. Taniuchi, H. Ito, K. Kawase, Enhancement of terahertz-wave output from LiNbO₃ optical parametric oscillators by cryogenic cooling. *Opt. Lett.* **24**, 202–204 (1999)
37. C. Janke, M. Forst, H. Nagel et al., Asynchronous optical sampling for high-speed characterization of integrated resonant terahertz sensors. *Opt. Lett.* **30**(11), 1405–1407 (2005)
38. S. Preu, G.H. Dohler, S. Malzer, L.J. Wang, A.C. Gossard, Tunable, continuous-wave terahertz photomixer sources and applications. *J. Appl. Phys.* **109**(6), 4 (2011)
39. N. Kim, J. Shin, E. Sim et al., Monolithic dual-mode distributed feedback semiconductor laser for tunable continuous-wave terahertz generation. *Opt. Express* **17**(16), 13851–13859 (2009)
40. S. Kumar, Recent progress in terahertz quantum cascade lasers. *IEEE J. Sel. Top. Quantum Electron.* **17**(1), 38–47 (2011)
41. M. Fischer, G. Scalari, C. Walther, J. Faist, Terahertz quantum cascade lasers based on In_{0.53}Ga_{0.47}As/In_{0.52}Al_{0.48}As/InP. *J. Crystal Growth* **311**(7):1939–1943 (2009)
42. M.Y. Frankel, R.D. Esman, Optical single sideband suppressed carrier modulator for wide-band signal processing. *J. Lightwave Technol.* **16**(5), 859 (1998)
43. P. Gu, F. Chang, M. Tani, K. Sakai, C.L. Pan, Generation of coherent cw-terahertz radiation using a tunable dual-wavelength external cavity laser diode. *Jpn. J. Appl. Phys.* **38**(11A), L1246 (1999)
44. N. Kim, S.P. Han, H. Ko, et al., Tunable continuous-wave terahertz generation/detection with compact 1.55 μm detuned dual-mode laser diode and InGaAs based photomixer. *Optics Expr.* **19**(16):13851–13859 (2011)

45. O. Morikawa, M. Tonouchi, M. Tani, K. Sakai, M. Hangyo, Sub-THz emissions properties of photoconductive antennas excited with multimode laser diode. *Jpn. J. Appl. Phys.* **38**(3R), 1388 (1999)
46. M.A. Belkin, F. Capasso, F. Xie et al., Room temperature terahertz quantum cascade laser source based on intracavity difference-frequency generation. *Appl. Phys. Lett.* **92**(20), 201101 (2008)
47. S. Mickan, D. Abbot, J. Munch et al., Analysis of system trade-offs for terahertz imaging. *Microelectron. J.* **31**(7), 503–514 (2000)
48. J.T. Darrow, X.C. Zhang, D.H. Auston, Power scaling of large-aperture photoconducting antennas. *Appl. Phys. Lett.* **58**, 25 (1991)
49. J.T. Darrow, X.C. Zhang, D.H. Auston, J.D. Morse, Saturation properties of large-aperture photoconducting antennas. *IEEE J. Quantum Electron.* **28**(6), 1607 (1992)
50. F.W. Smith, A.R. Calawa, C.L. Chen et al., New MBE buffer used to eliminate backgating in GaAs MESFETs. *IEEE Electron Device Lett.* **9**(2), 77 (1988)
51. K.C. Zhang, D.H. Auston, Optoelectronic measurement of semiconductor surfaces and interfaces. *J. Appl. Phys.* **71**, 326 (1992)
52. T. Dekorsy, T. Pfeifer, W. Kutt, H. Kurz, Subpicosecond carrier transport in GaAs surface-space-charge fields. *Phys. Rev. B.* **47**, 3842 (1993)
53. J.N. Heyman, P. Neocleous, D. Hebert, P.A. Crowell, T. Müller, K. Unterrainer, Terahertz emission from GaAs and InAs in a magnetic field. *Phys. Rev. B.* **64**:085202 (2001)
54. H. Hamster, A. Sullivan, S. Gordon, W. White, R.W. Falcone, Subpicosecond, electromagnetic pulses from intense laser-plasma interaction. *Phys. Rev. Lett.* **71**, 2725 (1993)
55. W.P. Leemans, C.G.R. Geddes, J. Faure et al., Observation of terahertz emission from a laser-plasma accelerated electron bunch crossing a plasma-vacuum boundary. *Phys. Rev. Lett.* **91**, 074802 (2003)
56. A.M. Bystrov, N.V. Vvedenskii, V.B. Gildenburg, Generation of terahertz radiation upon the optical breakdown of a gas. *JETP Lett.* **82**, 753 (2005)
57. D.J. Cook, R.M. Hochstrasser, Intense terahertz pulses by four-wave rectification in air. *Opt. Lett.* **25**(16), 1210–1212 (2000)
58. X. Xie, J. Dai, X.C. Zhang, Coherent control of THz wave generation in ambient air. *Phys. Rev. Lett.* **96**(7), 75005 (2006)
59. J. Dai, X. Xie, X.C. Zhang, Detection of broadband terahertz waves with a laser-induced plasma in gases. *Phys. Rev. Lett.* **97**, 103903 (2006)
60. J. Dai, Laser air photonics: beyond the terahertz gap. *Mater. Today* **15**, 52 (2012)
61. K. Wynne, J.J. Carey, An integrated description of terahertz generation through optical rectification, charge transfer, and current surge. *Opt. Commun.* **256**, 400 (2005)
62. J.J. Carey, R.T. Bailey, D. Pugh et al., Terahertz pulse generation in an organic crystal by optical rectification and resonant excitation of molecular charge transfer. *Appl. Phys. Lett.* **81**, 4335 (2002)
63. J. Xu, A.B. Stickrath, P. Bhattacharya et al., Direct measurement of the photoelectric response time of bacteriorhodopsin via electro-optic sampling. *Biophys. J.* **85**(2), 1128 (2003)
64. Y.R. Shen, *The principles of nonlinear optics* (Wiley, New York, 1984), p. 110
65. G.K. Kitaeva, Terahertz generation by means of optical lasers. *Laser Phys. Lett.* **5**(8), 568 (2008)
66. J.A. Armstrong, N. Bloembergen, J. Ducuing et al., Interactions between light waves in a nonlinear dielectric. *Phys. Rev.* **127**, 1918 (1962)
67. M.M. Fejer, G.A. Magel, D.H. Jundt et al., Quasi-phase-matched second harmonic generation: tuning and tolerances. *IEEE J. Quantum Electron.* **28**(11), 2631 (1992)
68. I. Tomita, H. Suzuki, H. Ito, et al., Terahertz-wave generation from quasi-phase-matched GaP for 1.55 μm pumping. *Appl. Phys. Lett.* **88**, 071118 (2006)
69. H.J. Bakker, S. Hunsche, H. Kurz, Investigation of anharmonic lattice vibrations with coherent phonon polaritons. *Phys Rev B* **50**, 914 (1994)
70. J.A. L'huillier, G. Torosyan, M. Theuer, et al., Generation of THz radiation using bulk, periodically and aperiodically poled lithium niobate—part 1: theory. *Appl. Phys. B.* **86**(2), 185 (2007)

71. J.A. L'huillier, G. Torosyan, M. Theuer, et al., Generation of THz radiation using bulk, periodically and aperiodically poled lithium niobate—part 2: experiments. *Appl. Phys. B.* **86**(2), 185–197 (2007)
72. A.K. Sharma, R.K. Patidar, M. Raghuramaiah et al., Measuring pulse-front tilt in ultrashort pulse laser beams without ambiguity of its sign using single-shot tilted pulse-front autocorrelator. *Opt. Expr.* **14**, 13131 (2006)
73. J. Hebling, G. Almasi, I. Kozma et al., Velocity matching by pulse front tilting for large-area THz-pulse generation. *Opt. Expr.* **10**(21), 1161–1166 (2002)
74. B. Bartal, I.Z. Kozma, A.G. Stepanov et al., Toward generation of μ J range sub-ps THz pulses by optical rectification. *Appl. Phys. B* **86**, 419 (2007)
75. Y.S. Lee, T. Meade, M. DeCamp et al., Temperature dependence of narrow-band terahertz generation from periodically poled lithium niobite. *Appl. Phys. Lett.* **77**(9), 1244 (2000)
76. Y.S. Lee, T. Meade, T.B. Norris et al., Tunable narrow-band terahertz generation from periodically poled lithium niobate. *Appl. Phys. Lett.* **78**(23), 3583 (2001)
77. K.L. Vodopyanov, M.M. Fejer, X. Yu et al., Generation of multicycle terahertz pulses via optical rectification in periodically inverted GaAs structures. *Appl. Phys. Lett.* **89**, 141119 (2006)
78. G.K. Kitaeva, K.A. Kuznetsov, V.F. Morozova et al., Reduction-induced polarons and optical response of Mg-doped LiNbO₃ crystals. *Appl. Phys. B* **78**, 759 (2004)
79. M.R.K. Soltanian, I.S. Amiri, S.E. Alavi et al., Dual-wavelength erbium-doped fiber laser to generate terahertz radiation using photonic crystal fiber. *J. Lightwave Technol.* **33**(24), 5038 (2015)
80. L. Dong, F. Xie, S. Ma et al., Simple tunable dual-wavelength fiber laser and multiple self-mixing interferometry to large step height measurement. *Opt. Express* **24**(19), 21880–21885 (2016)
81. S. Pan, X. Zhao, C. Lou et al., Switchable single-longitudinal-mode dual-wavelength erbium-doped fiber ring laser incorporating a semiconductor optical amplifier. *Opt. Lett.* **33**(8), 764–766 (2008)
82. X. Fan, W. Zhou, S. Wang et al., Compact dual-wavelength thulium-doped fiber laser employing a double-ring filter. *Appl. Opt.* **55**(12), 3319–3322 (2016)
83. H. Ahmad, A.A. Latif, M.Z. Zulkifli et al., High power dual-wavelength tunable fiber laser in linear and ring cavity configurations. *Chin. Opt. Lett.* **10**(1), 010603 (2012)
84. A.J. Deninger, A. Roggenbuck, S. Schindler, et al., 2.75 THz tuning with a triple-DFB laser system at 1550 nm and InGaAs photomixers. *J. Infrared, Millimeter, Terahertz Waves* **36**(3), 269–277 (2015)
85. M.A. Umyy, S. Bikorimana, R. Dorsinville, Beam Combining of SOA-based bidirectional tunable fiber compound-ring lasers with external reflectors, in *Paper presented at the 5th international conference on photonics, optics, and laser technology*, Porto, Portugal, 27–31 Feb 2017
86. H. Bruesselbach, D.C. Jones, M.S. Mangir et al., Self-organized coherence in fiber laser arrays. *Opt. Lett.* **30**(11), 1339–1341 (2005)
87. B. Globisch, R.J.B. Dietz, S. Nellen et al., Terahertz detectors from Be-doped low-temperature grown InGaAs/InAlAs: interplay of annealing and terahertz performance. *AIP Adv.* **6**(12), 125011 (2016)
88. G. Carpintero, E. Garcia-Munoz, H. Hartnagel et al., *Semiconductor Terahertz Technology: Devices and Systems at Room Temperature Operation* (Wiley, New York, 2015)

Chapter 7

Reliability Challenges of Nanoscale Avalanche Photodiodes for High-Speed Fiber-Optic Communications



Jack Jia-Sheng Huang, Yu-Heng Jan, H. S. Chang, C. J. Ni, Emin Chou, S. K. Lee, H. S. Chen and Jin-Wei Shi

Abstract Photodetectors in optical systems work in a similar manner like human eyes. Optical detectors can detect signals from light sources and provide feedback to the networks. Modern nanoscale semiconductor photodetectors are indispensable components for various high-speed optical networks in the applications of datacenter, wireless, fiber-to-the-premises, and telecommunication. In this chapter, we focus on the state-of-the-art 2.5G, 10G, and 25G avalanche photodiodes and compare the feature size in each generation. We present brief overview of the key device performance of avalanche photodiodes including avalanche breakdown voltage, dark current, temperature stability, bandwidth, and sensitivity. We also discuss reliability implications associated with device miniaturization. During device shrinking, increasingly high electric field is likely to impose most reliability risk. We discuss the reliability challenges of nanoscale photodetectors in terms of optical/electrical overload stress, wear-out degradation, and electrostatic discharge.

7.1 Introduction

The functions of photodetectors for optical networks are comparable to how eyes work for humans. As a sense organ, the human eye allows vision and helps provide a three dimensional, moving image, normally colored in daylight [1]. Similarly, photodetectors are sensors of light that are vital to communicate with laser sources in

J. J.-S. Huang (✉) · Y.-H. Jan
Source Photonics, 8521 Fallbrook Avenue, Suite 200, West Hills, CA 91304, USA
e-mail: jack.huang@sourcephotonics.com

J. J.-S. Huang · Y.-H. Jan · H. S. Chang · C. J. Ni · E. Chou · S. K. Lee · H. S. Chen
Source Photonics, No.46, Park Avenue 2nd Rd., Science-Based Industrial Park, Hsinchu, Taiwan

J.-W. Shi
Department of Electrical Engineering, National Central University, Zhongli, Taiwan

© Springer Nature Switzerland AG 2019
P. Ribeiro and M. Raposo (eds.), *Optics, Photonics and Laser Technology 2018*, Springer Series in Optical Sciences 223,
https://doi.org/10.1007/978-3-030-30113-2_7

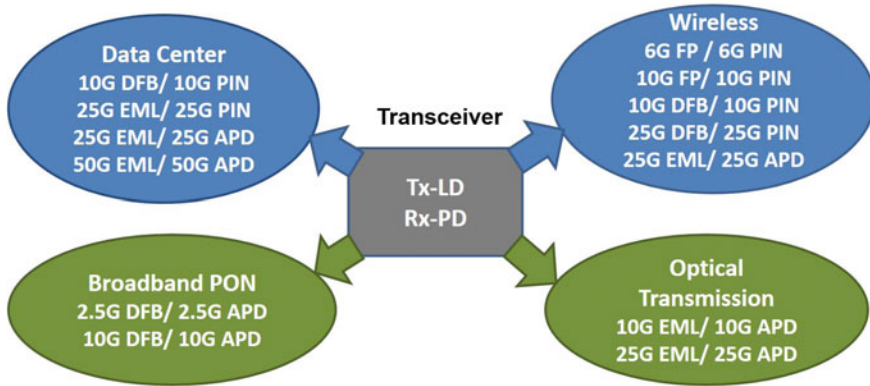


Fig. 7.1 Schematics of main optical segments and their respective semiconductor components

fiber optic networks [2]. A photodetector has a p–n junction that converts light photons into current [3]. The absorbed photons make electron–hole pairs in the depletion region. Photodiodes and phototransistors are a few examples of photodetectors.

There are two main types of photodiodes: p-type/intrinsic/n-type photodiodes (PIN PDs) and avalanche photodiodes (APDs) [4]. A PIN PD is a diode with a wide, undoped intrinsic semiconductor region between a p-type semiconductor and an n-type semiconductor region. The intrinsic region absorbs the light and convert it to electrical signal. The p-type and n-type regions are typically heavily doped because they are used to make ohmic contacts. On the other hand, APDs operate based on the principle of semiconductor avalanche breakdown. The APDs have drawn increasing interest in the commercial and military applications due to their performance advantages of high-bandwidth, high sensitivity, and low noise [5–10].

Photodiodes are widely deployed in various fiber optic applications [11]. Figure 7.1 shows the four key areas of fiber optic communications. The first is data-center where a new wave of rapid growth has been fueled by the consumer demand for higher internet bandwidth. The second is wireless where the next generation is poised to penetrate to each corner of our lives. The third is passive optical network (PON), a technology that enables high-speed data access for fiber-to-the-premises or fiber-to-the-homes. The fourth is optical transmission where long-haul telecommunication networks carry huge loads of information between cities, through mountains, and from coast to coast.

7.1.1 Datacenter Network

Figure 7.2 shows illustration of datacenter networks that empower companies such as Google, Facebook, Microsoft, and Amazon. The common building blocks for the light sources include 10G distributed feedback (DFB) laser, 25G electro-absorption

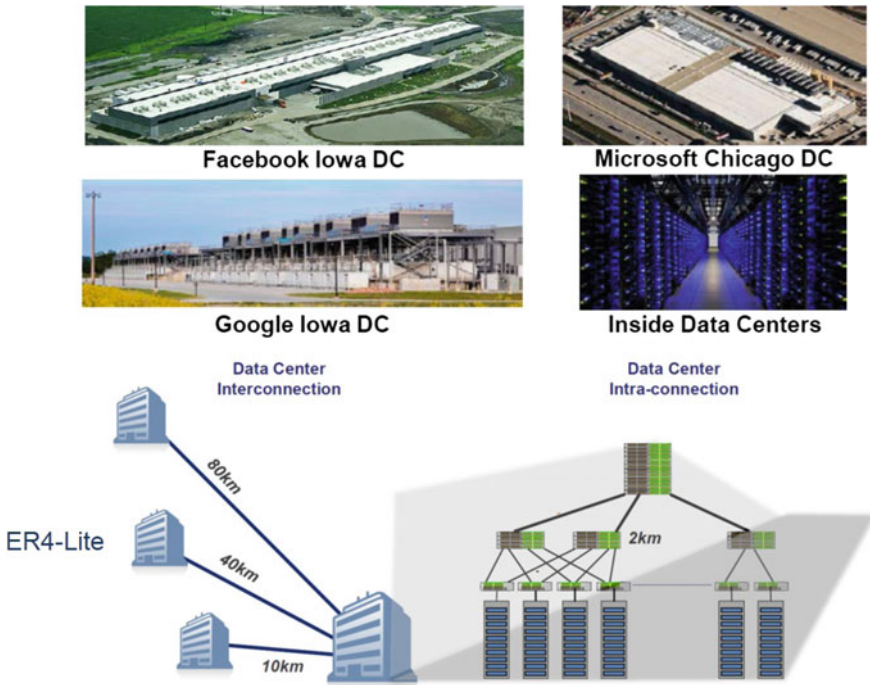


Fig. 7.2 Schematics of datacenter networks that empower mega datacenters for companies such as Google, Facebook, Microsoft, and Amazon

modulated laser (EML), and 25G directly modulated lasers (DML). The light detectors include 10G PIN PD and 25G PIN PD. For example, the modern 100G QSFP transceiver is comprised of the 25G EML and 25G PIN. Most recently, 25G APD has been demonstrated in 40 km fiber link for ER4-Lite network [12]. Table 7.1 shows the 40/100G high-speed ethernet for various optical links. Historically, the datacenter communications was limited to 10 km in LR4. In recent years, ER4 applications have

Table 7.1 IEEE 802.3 standards for 40G/100G optical modules and fiber optics [12]

40G/100G ethernet	Transmission distance	Optical fiber
SR4	125 m	Multi-mode fiber (MMF)
FR4 (CWDM4)	2 km	Single-mode fiber (SMF)
LR4	10 km	SMF
ER4 Lite /ER4	30 km /40 km	SMF

gained significant tractions because of the technological advantages in the 30–40 km optical reach.

7.1.2 Wireless Network

For wireless, a wide spectrum of components are employed to cover the applications of third generation (3G), fourth generation (4G), and fifth generation (5G) mobile networks [13]. Wireless light sources include 6G Fabry-Perot (FP), 10G FP, 6G DFB, 10G DFB, 25G DFB, and 25G EML [11]. Wireless photodetectors include 6G PIN PD, 10G PIN PD, 25G PIN PD, and 25G APD.

Figure 7.3 illustrates the schematics of 5G mobile network where wireless signals provide optical connection to automobile, drone, artificial intelligence machine, stadium, aircraft, satellite, etc. [14]. 5G wireless offers speeds about 100 times than 4G. The unprecedented high-speed of 5G enables the connectivity over a large scope of objects including self-driving vehicle, internet of things (IoT), drones, robots, and other artificial intelligence (AI) machines.

The 2018 Winter Olympic Games in PyeongChang, South Korea debuted the first large-scale public demonstrations of 5G video streaming technologies, broadcasting, as well as the world's first major 8 K high-definition video production [15]. The 5G-linked cameras allowed viewers to watch sporting events from all angles in real time.

For the future, 5G is expected to play a very critical role in the Tokyo 2020 Olympics [16, 17]. Japan is working with NTT to trial 5G long-distance and high-speed data transmission between Tokyo SkyTree and surrounding shopping centers.

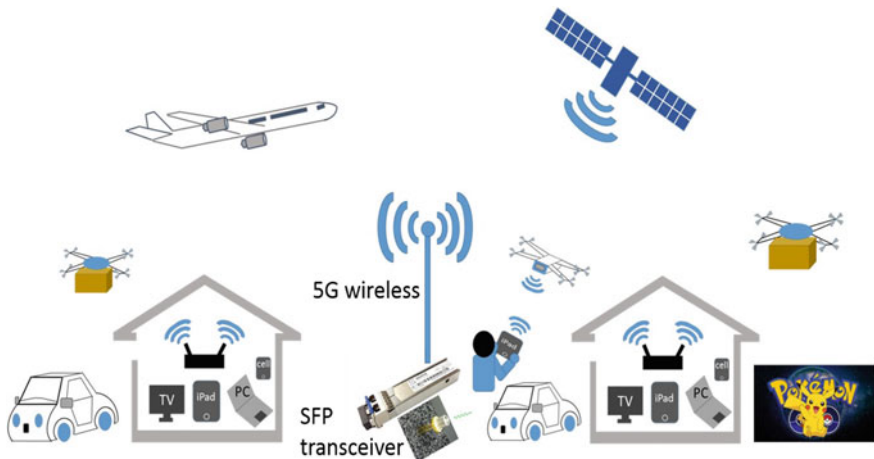


Fig. 7.3 Schematics of fifth generation (5G) wireless network where wireless signals provide connectivity to automobile, drone, artificial intelligence, stadium, aircraft, satellite, etc. [11]

7.1.3 Passive Optical Network

Figure 7.4 shows the schematics of PON architecture where the optical line terminal (OLT) at the service provider's central office connects with a number of optical network units (ONUs) at multiple points near the end users [18, 19]. The upstream and downstream signals are transmitted between the OLT and ONU through optical distribution network (ODN) that consists of the fibers and splitters between them. For both OLT and ONU, transmitter and receiver are the main active components to send and receive the optical signals.

For ONU, the light sources for transmitters are mainly based on 2.5G DFB and 10G DFB, and the detectors for receivers typically comprise 2.5G APD and 10G APD. For OLT, 10G EML and 10G APD are commonly deployed in transmitters and receivers, respectively.

7.1.4 Telecommunication Network

For telecommunication, key components include 10G EML and 10G APD. Telecommunication applications typically require long-distance transmissions in the range of 40–80 km. The EML emitter with lasing wavelength near 1550 nm is usually employed to achieve low attenuation over fiber [20–23]. On the detector side, 10G APD is widely deployed because of its performance edges in high sensitivity, high bandwidth, and low noise [6, 7, 24, 25].

Figure 7.5 shows telecommunication network where switched networks span large distances, such as interstate and international communications. Long-haul telecommunication links carry huge loads of information between cities, through mountains, and from coast to coast. Within the city or metropolitan area, short-haul fibers connect cellular towers, enterprise buildings, datacenters, and homes with intermediate access

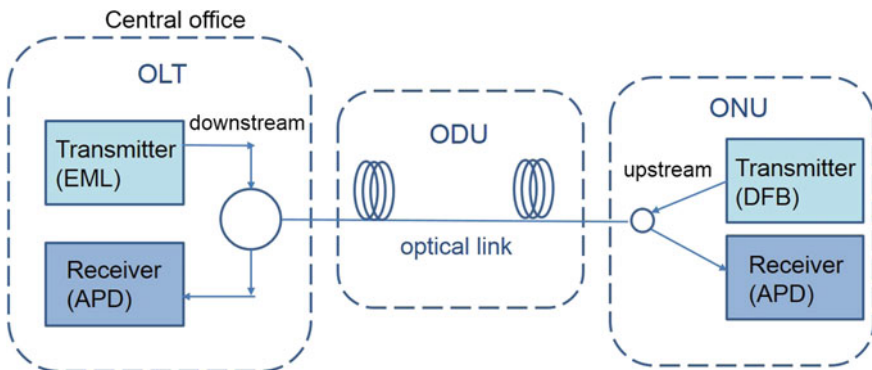


Fig. 7.4 Schematics of PON where OLT at the service provider's central office hub communicate with a number of ONUs near the end users

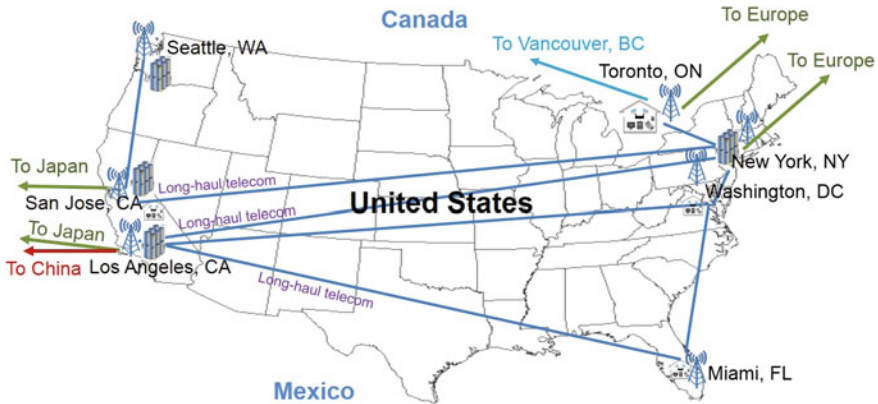


Fig. 7.5 Schematics of telecommunication network where long-haul optical links carry huge loads of information between cities, through mountains, and from coast to coast. Within the city or metropolitan area, short-haul fibers connect cellular towers, enterprise buildings, datacenters, and homes with intermediate access points

points. Businesses use telecommunications to expand and grow their networks. With computer, internet, and telephone networks, businesses can allocate their resources efficiently. For example, the computer network consists of computers and devices connected to one another. Computer networks can range from a local area network (LAN) to a wide area network (WAN). The internet network enables users to discover information almost instantly and apply concepts to different situations. The telephone network is used to route calls as well as to service their customers.

7.2 APD Light Detectors

Table 7.2 lists some of the common APD components used in commercial and military applications [26]. In terms of volume and history, 2.5G APD is the most widely fabricated avalanche photodetector. The deployment of 2.5G APD is typically in PON systems. It's based on planar structure where P- and N-metal are on the top and bottom, respectively. The 2.5G APD device illustrated in Table 7.2 is comprised of p-InP/i-InP/n-InP/i-InGaAs/n-InP. The topmost p-type InP layer is formed by Zn diffusion [27]. Guard ring is normally added to avoid the edge breakdown at the p-InP diffused region. The second layer below the p-type InP is the i-InP (intrinsic) that serves as the multiplication layer (M-layer). This is the critical layer for APD since it provides gain multiplication function. Next, a thin n-type InP layer (~100 nm) is used as the charge control. The i-InGaAs layer (2 μm in this illustration) is used for light absorption. Finally, an n-type InP buffer layer is inserted between the i-InGaAs absorption layer and n-InP substrate.

Table 7.2 List of key commercial photodetector components including 2.5G APD, 10G APD, and 25G APD. The schematics are shown to indicate the absorption, charge control, and multiplication layers [26]

	Device structure	Key feature
2.5G APD	<p>The schematic shows a planar structure. At the top, there is a P-metal layer with an AR (anti-reflection) coating and a Zn diffused region forming a guard ring. Below this is a p-InP layer. The active region consists of an i-InP multiplication layer (500 nm thick), an n-InP charge control layer (100 nm thick), and an i-InGaAs absorption layer (2 μm thick). Below the absorption layer is an n-InP layer (1 μm thick), and at the bottom is an n-InP sub. layer. The entire structure is on an N-metal substrate.</p>	<ul style="list-style-type: none"> - Planar structure - P-metal on top - N-metal on bottom - InP multiplication layer (~500 nm) formed by Zn diffusion - Charge control layer (~100 nm) by epitaxial growth
10G APD	<p>The schematic shows a mesa structure. From top to bottom: P-metal, SiNx AR, Dielectric, InGaAs contact, InP window, Graded layer, InGaAs absorption layer (1.2 μm thick), Graded layer, InAlAs charge control layer (50 nm thick), InAlAs multiplication layer (200 nm thick), N-InP contact (800 nm thick), N-InP buffer, and S.I.-InP substrate. P-mesa and N-mesa are indicated on the right side.</p>	<ul style="list-style-type: none"> - Mesa structure - P- and N-metal on top - Increasing diameter of mesa from top to bottom - InAlAs multiplication layer (200 nm) grown by MBE - Charge control layer (50 nm) grown by MBE
25G APD	<p>The schematic shows a mesa structure. From top to bottom: P-metal, SiNx AR, Dielectric, InGaAs contact, InP window, InAlGaAs graded, InGaAs absorption (430 nm thick), InGaAs absorption layer (370 nm thick), InAlGaAs graded, InAlAs charge control layer (30 nm thick), InAlAs field buffer layer (60 nm thick), InAlAs charge control layer (30 nm thick), InAlAs multiplication layer (88 nm thick), N-InAlAs contact (200 nm thick), N-InP contact (1000 nm thick), N-InP buffer, and S.I.-InP substrate. P-mesa and N-mesa are indicated on the right side.</p>	<ul style="list-style-type: none"> - Mesa structure - P- and N-metal on top - Increasing diameter of mesa from top to bottom - Smaller diameter - Dual charge layer - InAlAs multiplication layer (100 nm) grown by MBE - Charge control layer (30 nm) grown by MBE

10G APD has also been gaining popularity in optical communication applications, particularly in 10G PON. The 10G APD offers several performance advantages such as high sensitivity, high signal-to-noise ratio, and low dark current [6]. It is based on mesa structure with coplanar P- and N-metal contacts to enhance the speed and reduce the series resistance [28]. The active region is sandwiched between the P-mesa at the top and the N-mesa at the bottom. The active region in this case consists of the InGaAs absorption (1.2 μm), InAlAs charge control (50 nm), and InAlAs multiplication (200 nm) layers. Such active structure is also called separate absorption, charge, and multiplication (SACM). These layers are formed in opposite order compared to 2.5G APD. Electron carrier is utilized for multiplication of 10G APD while hole carrier is used for the multiplication of 2.5G APD [26]. The P-mesa is made up of p-type InP window and p-type InGaAs contact layers. The p-contact is usually in the form of metal ring that surrounds the circular anti-reflective (AR) window. The p-metal ring is connected to the outside p-metal pad by a metal bridge. The N-mesa at the bottom consists of the N-InP buffer and contact layers, grown on a semi-insulating (S.I.) InP substrate. The n-metal contact is connected to the N-mesa. For the passivation, low-k dielectric material such as polyimide is used to reduce the capacitance.

For 25G APD, mesa structures with more complex flavors are designed to achieve the ever-increasing speed requirement. As of today, 25G APD is mostly available for prototype samples. High-volume manufacturing with robust reliability performance remains challenging. 25G APD shares some commonality in structure compared to 10G APD. For example, both 10G and 25G APDs are based on SACM structures that operate on electron carrier multiplication. The order of absorption, charge control, and multiplication layers from top to bottom is generally the same. On the other hand, the main differences between 10G and 25G APDs include (1) diameter, (2) absorption layer thickness, and (3) charge layer thickness and design. For instance, the diameter of 25G APD is about 26–30 μm , while that of 10G APD is typically > 40 μm . For device scaling, the thicknesses of the critical regions such as absorption, charge control, and multiplication layer are accordingly reduced to attain higher speed. In addition, dual charge layers are sometimes employed to reduce the electric breakdown [7].

7.2.1 Reverse IV and Dark Current

Figure 7.6a shows the typical reverse current-voltage (IV) curve of a 10G mesa-type SACM APD photodetector. The IV curve shows two transitions at about 10 V and 32 V, respectively. The first transition at 10 V corresponds to the punch-through voltage at which electric field depletes both the InAlAs multiplication and *i*-InGaAs absorption layers [6, 29]. The punch-through voltage related to the absorption layer is denoted as V_{pt} . The second transition represents the avalanche breakdown voltage (V_{br}) that is typically determined by the InAlAs multiplication layer. The breakdown voltage is typically taken at 10 μA on the IV curve. The dark current at 0.9 V_{br} is estimated to be about 12 nA at 25 °C. In the case of 25G APD shown in Fig. 7.6b,

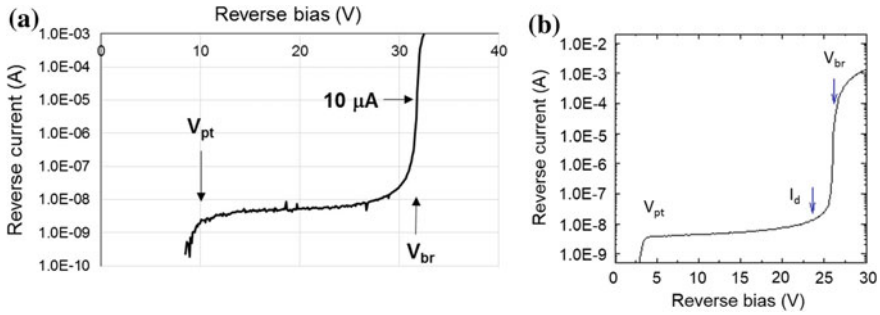


Fig. 7.6 The reverse IV curves of **a** 10G APD and **b** 25G APD devices measured at 25 °C. Both 10G and 25G APDs are mesa structures [6, 12]

the two transitions also occur on the reverse IV, but at lower voltages compared to those of 10G APD. The punch-through voltage and avalanche breakdown voltage are about 3 V and 26 V, respectively. The dark current of 25G APD is estimated to be about 13 nA at 25 °C.

7.2.2 Electric Field Control

The APD device operates at avalanche breakdown regime at very high electric field ($\sim 10^5$ to 10^6 V/cm). Figure 7.7 shows the internal electric field schematic of a typical SACM APD [30]. The electric field of the absorption layer is typically low. In the charge control layer, electric field starts to build up. The electric field reaches the maximum in the multiplication layer that is responsible for avalanche breakdown operation.

High electric field at the multiplication layer is of concern owing to edge breakdown issue [7]. Several remedy designs have been experimented to circumvent the breakdown issue. For example, Nada et al. employed an inverted InAlAs APD structure [8]. However, such inverted APD may impose design constraint in the charge and passivation layers due to the concerns of surface breakdown. More recently, Chen et al. demonstrated a novel N-side down InAlAs APD with dual charge layers. By etching a mesa structure through the extra charge layer above the multiplication layer, the electric field can be strongly confined within the center of the multiplication layer [7]. As shown in Fig. 7.8a, the first mesa with a 30 μm diameter, is etched through the upper charge layer and stops at the 60 nm $\text{In}_{0.52}\text{Al}_{0.48}\text{As}$ buffer layer. The additional charge layer in the upper mesa is expected to effectively confine the strong electric field at the bottom multiplication layer within the range of 30 μm diameter. Figure 7.8b, c show the simulated electric-field distributions in vertical and horizontal directions near breakdown voltage (-16.5 V). The dopings in the two charge layers have been further optimized to avoid the problem of absorption layer breakdown that may limit the gain-bandwidth product of APD. Furthermore, as shown in Fig. 7.8c,

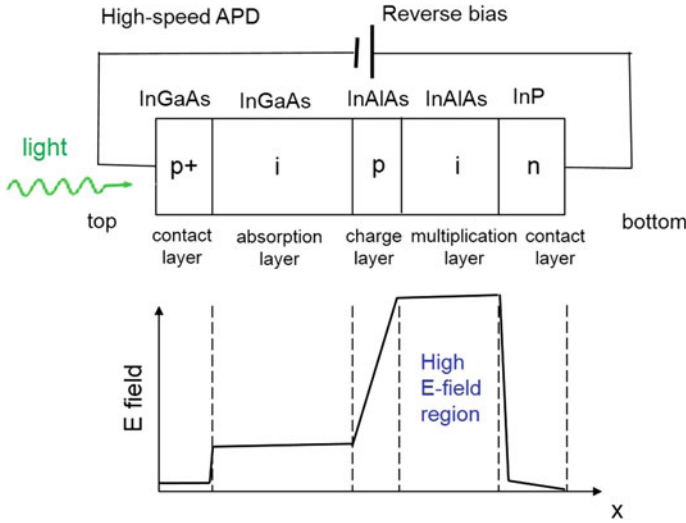


Fig. 7.7 Internal electric-field profile of the SACM APD structure. In the illustration, the InGaAs absorption layer is on top of the InAlAs charge control and multiplication layers [30]

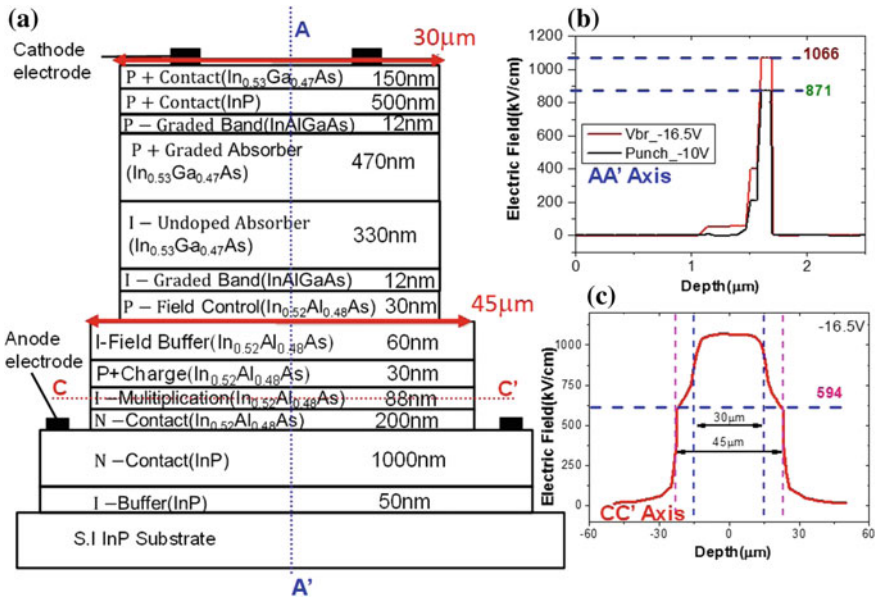


Fig. 7.8 a Conceptual cross-sectional view of the demonstrated device. The simulated electric-field distributions in b vertical (A-A') and c horizontal directions (C-C') [7]

the horizontal electric field in the multiplication layer is well confined in the 30 μm diameter (1000 kV/cm), and the electric field at the edge of second mesa can be greatly reduced to around 590 kV/cm to suppress the edge breakdown.

7.2.3 Temperature Stability

There are two major types of junction breakdown in semiconductor diodes. The first type is associated with the tunneling breakdown where the tunneling mechanism is dominant for the devices with lower breakdown voltage. The tunneling breakdown process exhibits a negative temperature coefficient [31, 32]. The second type is the avalanche breakdown that is the dominant mechanism for large breakdown voltage. The temperature coefficient of the avalanche breakdown is positive.

In this chapter, we focus on the avalanche breakdown as illustrated in Fig. 7.9a. The avalanche breakdown voltage of APD is expected to increase due to the effect of phonon [33]. As the temperature increases, the population of phonons increases. Thus, a higher electric field and applied voltage are required to reach breakdown in order to overcome the increased carrier cooling caused by phonon scattering [34]. Figure 7.9b shows the normalized avalanche breakdown voltage of APD as a function of temperature. By extracting from the slope of Fig. 7.9b, the normalized temperature coefficient (α) can be deduced to be $5.1 \times 10^{-4} \text{ }^\circ\text{C}^{-1}$, which is better than the previously reported value of $7.2 \times 10^{-4} \text{ }^\circ\text{C}^{-1}$ [35]. The superior temperature stability of the mesa-type InAlAs APD can be attributed to the optimized design of the InAlAs multiplication layer thickness and InAlAs charge control layer doping.

The positive temperature coefficient of the APD was empirically instituted by Tyagi [36]. Assuming a linear relationship, the avalanche breakdown voltage can be described in terms of temperature as shown in (7.1).

$$V_{br}(T) = V_{br}(T_0)[1 + \alpha(T - T_0)] \tag{7.1}$$

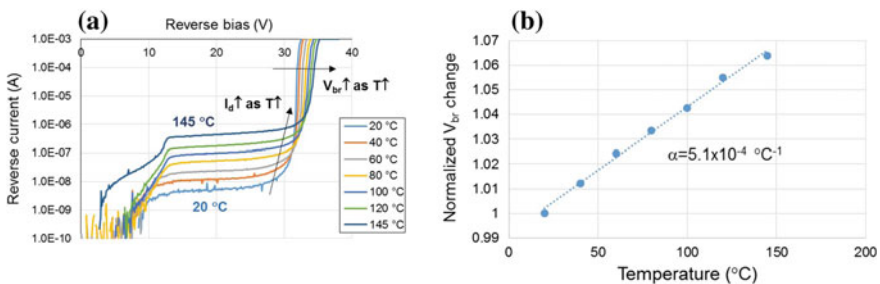


Fig. 7.9 **a** The reverse IV curve of a SCAM APD as a function of temperature and **b** normalized breakdown voltage versus temperature showing the positive temperature coefficient [6]

where $V_{br}(T)$ is the avalanche breakdown voltage at temperature T , $V_{br}(T_0)$ is the avalanche breakdown voltage at reference temperature T_0 , and α is the normalized temperature coefficient.

7.2.4 Bandwidth

Figure 7.10 shows the frequency response (bandwidth) plot of 10G and 25G APD devices based on small-signal modulation response at 25 °C. The 3 dB bandwidths of 10G and 25G APDs are typically around 7 and 17 GHz, respectively.

7.2.5 Sensitivity

Figure 7.11 shows the sensitivity plots of 10G and 25G APD devices. The sensitivity is estimated by the bit error rate (BER). The sensitivity typically decreases with increasing device speed. For 25G APD, the sensitivity of -21 dBm is sufficient to meet the ER4-Lite requirement. Such sensitivity level can provide good margin for photoreceiver detection over the 40 km transmission over fiber [37, 38].

Fig. 7.10 Frequency response plot of 10G and 25G APD photodetectors. The 3 dB bandwidth can be determined from the plot

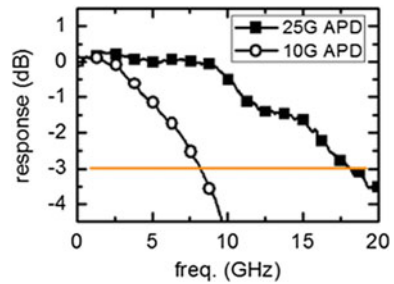
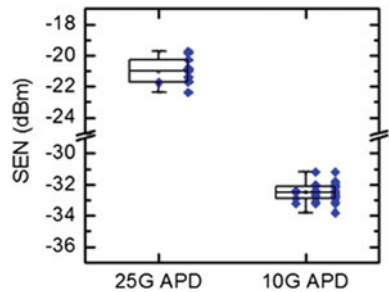


Fig. 7.11 Sensitivity plot of 10G and 25G APD devices. The sensitivity is based on bit error rate measurement



7.3 Device Miniaturization

Device miniaturization is the governing rule that drives the increased functionality and lower cost for semiconductors. However, each generation of device shrinking may also impose potential reliability risk that needs to be carefully evaluated. Figure 7.12 shows the timelines of photonic versus electronic device failure modes. For each technology, the main failure mechanism is highlighted where the proven device failures are marked in solid while the the new failure modes that involve forward-looking predictions or ongoing investigations are marked in dashed lines.

In the 1990s, the reliability failure modes of the electronic IC devices were primarily dominated by stress-induced void and contact/via electromigration when the technology nodes evolved from 0.9 μm to 0.18 μm [39–41]. Since early 2000, the failure mechanism related to Cu metallization has become important for 0.13 μm technology and beyond [42, 43]. For example, copper electromigration was widely observed from 65 nm to 32 nm nodes [42]. As the gate oxide became as thin as a few atomic monolayers, time-dependent dielectric breakdown (TDDB) failure related to low-k dielectric breakdown emerged to be one of the most challenging IC reliability issues for 16 nm technology [44, 45]. Nowadays, IC manufacturers have invested more R&D efforts in reliability studies related to solder packaging. As the solder bump shrinks further in the future, more reliability issues associated with higher current density and Joule heating are expected to occur [46, 47]. One of the modern semiconductor R&D avenues is 3-D IC that brings about new areas of reliability studies such as humidity and thermal cycling tests in the heterogeneous integration systems [48, 49].

For photodiodes (shown in orange), the common reliability issue in the 1990s was also related to Zn diffusion from the Au/Zn p-metallization. The p-metal reliability issue was largely mitigated after the adoption of the Ti/Pt/Au metal system. Starting from early 2000, dark current stability has become one of the key reliability issues for photodiodes [5, 25]. The dark current increase was particularly pronounced in

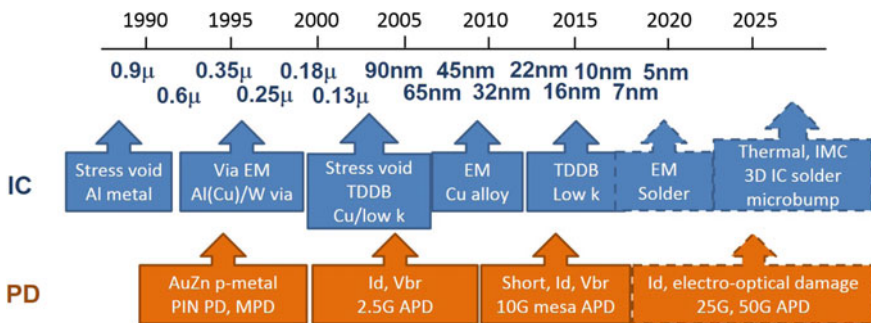


Fig. 7.12 The key reliability failure modes as a function of each technology node for electronic IC and photonic APD devices [11]

the APDs because of the higher electric field [6, 7]. The use of mesa structures in high-speed 10G and 25G APDs makes the stability of dark current and breakdown voltage even more challenging. In the mesa-type APD, dark current failure is often exacerbated by the presence of mesa surface defects and the high electric field [50, 51]. For 25G and beyond, mesa surface defects and electro-optical stress-induced degradation may become more pronounced as chip size and layer thickness continue to shrink.

Today, semiconductor devices are often required to incorporate design-in reliability in the early development phase [11, 52, 53]. During the device miniaturization, several factors including high electric field, Joule heating, and current crowding may impose reliability challenges.

(1) High electric field

It is well known that high electric field may induce reliability issues such as electromigration in IC interconnect [39–41, 54, 55]. Similarly, the high electric field of APD resulting from decreased thickness of multiplication layer may also impose reliability issue such as device breakdown or shorting [25, 56].

(2) Joule heating

As device size shrinks, Joule heating is expected to increase at a given bias current due to increased series resistance [57–59]. The increased Joule heating would raise the device junction temperature and in turn degrade functional performance and long-term reliability.

(3) Current crowding and non-uniform geometry

Localized non-uniform geometry may induce current crowding and cause early reliability failure at the weak site as manifested in the IC interconnects [60–63]. For APD, current crowding may also accelerate device degradation in both planar and mesa structures [26, 32].

7.4 Future Reliability Challenges

Among the aforementioned three factors, high electric field is expected to impose the most reliability challenge for APDs. This is because the effect of Joule heating can be mitigated by the use of mesa structure where the resistance of the mesa-type APD is typically lower than that of the planar-type due to the shorter conduction path [64]. The lower resistance would result in lower Joule heating. Regarding non-uniform geometry, the local surface states of mesa-type APD can be largely minimized by implementing careful processing control. For instance, the dark current of modern mesa-type APD devices can be controlled to within a few tens of nano ampere [6, 7].

Fig. 7.13 Evolution of minimum feature size for APD and IC. Device miniaturization occurs for both photonic and electronic devices to attain higher speed [26]

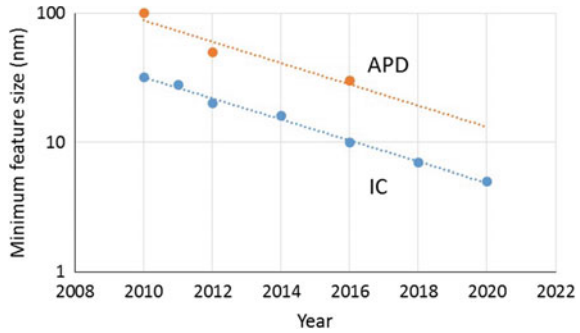
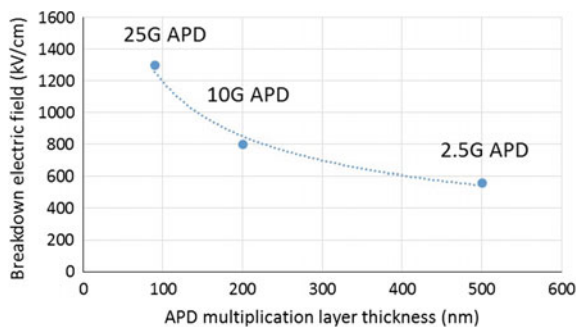


Figure 7.13 shows the evolution of the minimum feature size of IC and APD devices. Evidently, the charge control layer of APD shows similar device scaling compared to that of IC where the IC device scaling has rigorously followed Moore’s law for decades [65].

Figure 7.14 shows the breakdown electric field as a function of the multiplication layer thickness of commercial 2.5G, 10G, and 25G APDs. At the breakdown voltage for avalanche operation, majority of the voltage drop typically occurs at the absorption region and multiplication layer [32]. Due to the thin layer, the electric field is much higher in the multiplication layer as shown in the electric field profile in Fig. 7.7. Assuming the voltage drop splits equally in the absorption and the multiplication layers, the electric breakdown field at the multiplication layer is estimated to be about 560, 800, and 1300 kV/cm for 2.5G, 10G, and 25G APDs, respectively [66].

The projected trend of increasing breakdown electric field suggests that reliability issue may need to be carefully studied for 25G APD and beyond. With the potential new failure mechanisms, the future high-speed APD devices may draw another new frontier of reliability studies.

Fig. 7.14 The breakdown electric field as a function of the multiplication layer thickness for 2.5G, 10G, and 25G APD devices. In general, the breakdown field increases with increasing device speed



In the following, we discuss the effect of high electric field on optical/electrical overload, aging degradation, and ESD for APD devices.

7.4.1 Optical and Electrical Overload Stress

Overload stress has been commonly tested to evaluate photodetector’s robustness against simultaneous electrical and optical stresses. Fiber optic component manufacturers typically perform overload stress in burst and continuous modes, as illustrated in Fig. 7.15. To determine the damage threshold of overload, optical stress is ramped up from -4 to $+4$ dBm when APD is subjected to electrical stress at 2 V below the breakdown voltage (V_{br}).

With robust design and process, the APD devices can achieve good survival rate under the harsh overload stresses. In the burst mode, the optical stress of 1% pulsed duty cycle is applied to the APD for 60 s. Taking the $10G$ APD as an example, the damage threshold of overload under the pulse mode is typically around $+4$ dBm. In the continuous wave (CW) mode, the damage threshold can also sustain the stress level of $+0$ dBm for $10G$ APD.

It is noted that the overload threshold of APD generally scales with device shrinking. Table 7.3 shows the breakdown electric field and damage threshold of $2.5G$, $10G$, and $25G$ APD devices. For $2.5G$ APD, the device under CW overload stress

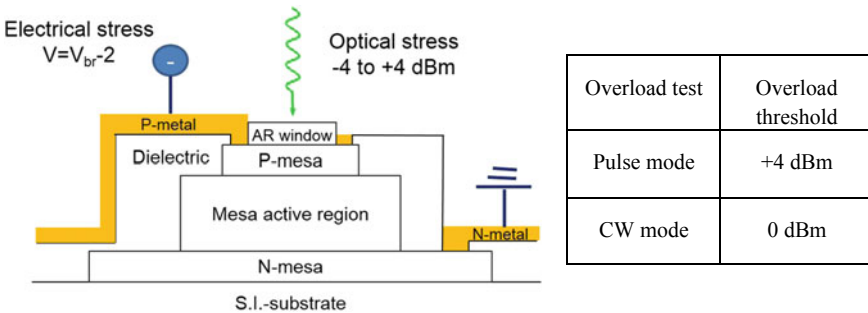


Fig. 7.15 Schematic of electrical and optical overload stress applied to the APD for both burst and CW modes [12]. The table shows the overload thresholds of $10G$ APD under pulse and CW modes

Table 7.3 Damage thresholds of optical and electrical overload stresses of $2.5G$, $10G$, and $25G$ APDs for burst and CW modes

Device	2.5G APD	10G APD	25G APD
Wavelength	1310 nm /1490 nm	1270 nm/1577 nm	1270–1330 nm
Breakdown electric field	560 kV/cm	800 kV/cm	1300 kV/cm
Damage threshold	+4 dBm (CW)	0 dBm (CW)	−2.5 dBm (CW)

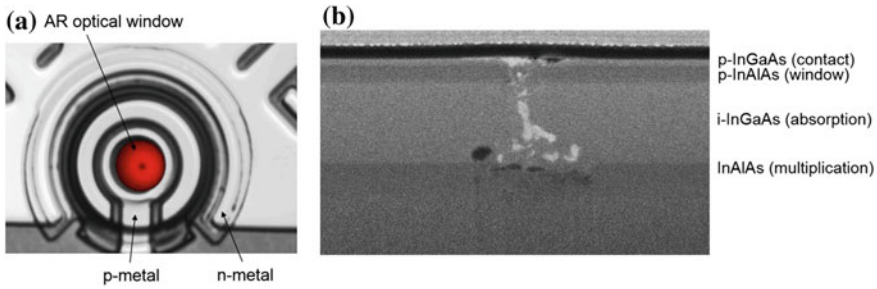


Fig. 7.16 **a** OBRICH image showing the thermal change spot. **b** Failure morphology of a damaged mesa-type 10G APD device after electrical and optical overload stress at +4 dBm pulse mode. The failure is related to shorting between the top InGaAs contact layer and the InAlAs multiplication layer

can typically sustain the damage threshold of +4 dBm. For 10G APD, the damage threshold of CW overload is decreased to 0 dBm due to the higher breakdown electric field. For 25G APD, the damage threshold of overload is further reduced to -2.5 dBm due to ultra-high breakdown field (~1300 kV/cm).

For the study of reliability physics, it is important to identify the failure location and morphology of APD device after the overload stress. Figure 7.16a shows the optical beam induced resistance change (OBIRCH) image where the area marked in red indicates the thermal change spot. Figure 7.16b shows an example of the damage morphology of a 10G mesa-type APD sample after overload stress of +4 dBm at burst mode. After the stress, the APD device shows functional failure due to short circuit near the InAlAs window. The short failure can be correlated with the metal shorting that shunts between the top InGaAs contact layer and the InAlAs multiplication layer, punching through the InGaAs absorption layer. Such short failure can be suppressed by employing an InP window structure. The failure morphology reveals the spatial location of the weak spot and provides a useful information about the robust overload design for the photodetector.

7.4.2 Reliability Aging

To ensure APD device robustness during field use, accelerated aging test is typically conducted in labs as the quick assessment of device reliability. The APD devices are tested with temperature and current stresses to study the degradation behavior [12].

The device failure time (t_f) follows the Black's equation [67–69] which provides a good empirical description of device degradation over time as a function of stress current density and temperature as shown in (7.2) [11, 12].

$$t_f = \frac{A}{j^N} \exp\left(\frac{E_a}{kT}\right) \quad (7.2)$$

In (7.2), the first term represents the current acceleration factor where A is a constant, j is the stress current density, and N is the current exponent; the second term represents the temperature acceleration where E_a is the activation energy, k is the Boltzmann's constant, and T is the temperature.

For the sake of activation energy study, (7.2) can be rewritten in the form of natural logarithm as shown in (7.3) where the third term can readily determine the activation energy.

$$\ln(MTTF) = \ln(A) - N \cdot \ln(j) + \frac{E_a}{kT} \quad (7.3)$$

7.4.2.1 Temperature Acceleration

Figure 7.17a shows the plot of $\ln(MTTF)$ versus $1/(kT)$ for 2.5G APD planar structures. The activation energy is extracted from the slope of the plot. Based on the experimental aging data of 240, 210 and 175 °C, the E_a is estimated to be 1.39 eV for the planar-type APD. The aging temperatures for the mesa APD structures are typically lower compared to the planar-type. Figure 7.17b shows the plot of $\ln(MTTF)$ vs. $1/(kT)$ for 10G APD mesa structures. Based on the experimental aging data of 165 and 175 °C, the E_a of the mesa-type APD is estimated to be 1.18 eV, in close agreement with other reported values [12]

Based on the E_a value, the failure times from the aging test can be extrapolated to project the device lifetimes at the operating condition by using (7.2). For the operating condition of 50 °C, the device lifetimes of the mesa-type 10G APD are estimated to be around 6900 years, which represents excellent reliability margin for

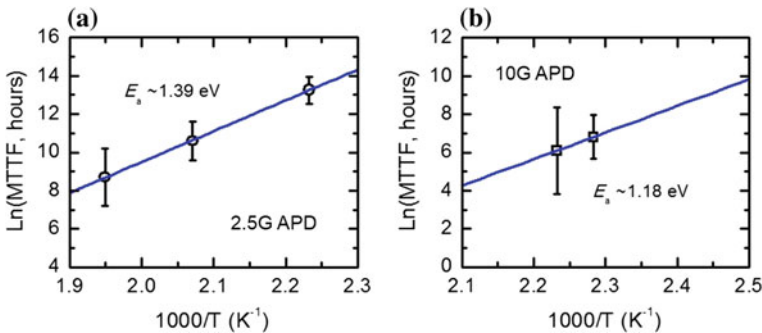
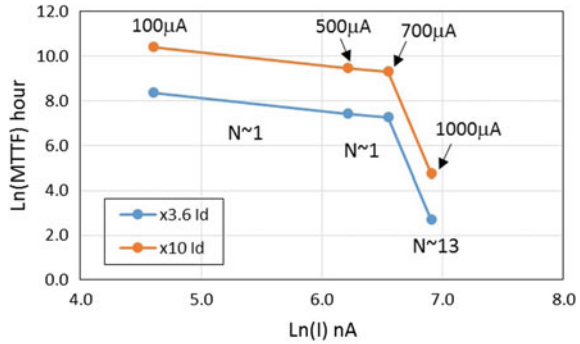


Fig. 7.17 The plot of $\ln(MTTF)$ versus $1/(kT)$ for **a** 2.5G APD planar structures and **b** 10G APD mesa structures. The E_a is taken from the slope according to (7.3). The E_a values are determined to be 1.39 and 1.18 eV for planar and mesa structures, respectively

Fig. 7.18 The plot of $\ln(\text{MTTF})$ versus $\ln(I)$ for the determination of N taken from mesa-type APD devices. The N value is 1 for the stress current of 100–700 μA , and N is increased to about 13 at the regime of overstress at 1000 μA likely due to thermal runaway



the 20-year stringent requirement per Telcordia. Knowledge of E_a would also allow one to quantitatively assess the influence of Joule heating in device lifetime. When Joule heating is small, the temperature term in (7.2) is roughly equivalent to ambient temperature or device junction temperature. When Joule heating is large, the junction temperature may need to be considered in Black’s equation as shown in (7.4) where ΔT is the temperature rise due to Joule heating.

$$t_f = \frac{A}{j^N} \exp \left[\frac{E_a}{k(T + \Delta T)} \right] \tag{7.4}$$

7.4.2.2 Current Acceleration

Figure 7.18 shows the plot of $\ln(\text{MTTF})$ versus stress current of mesa-type 10G APD devices. There are two regimes in the current acceleration. The first regime corresponds to $N \sim 1$ when the reverse stress current is in the range of 100–700 μA . The second regime shows large value of N (~ 13) at high stress current ($=1000 \mu\text{A}$) likely due to thermal runaway caused by excessive Joule heating.

7.4.3 Electrostatic Discharge

ESD is another important aspect of semiconductor device reliability [70, 71]. ESD damage could result from human operators, machine transient, current surge, etc. For APD, the thinner multiplication layer thickness, the higher the current density the ESD stresses may incur. This would lead to increased ESD sensitivity and reduced damage threshold. Table 7.4 compares the human-body-model (HBM) ESD thresholds of 2.5G and 10G APD devices. For forward bias, the ESD threshold of 10G APD is higher than that of 2.5G APD due to lower series resistance. For reverse bias,

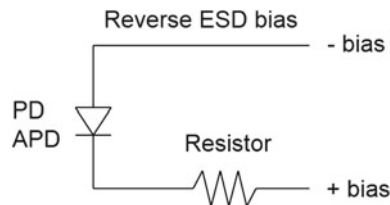
Table 7.4 Human-body-model ESD thresholds of 2.5G planar-type and 10G mesa-type APDs [64]

HBM ESD	2.5G APD (planar-type)	10G APD (mesa-type)
Forward bias	+2750 V	+5500 V
Reverse bias	−500 V	−150 V

the ESD threshold value of 10G APD is lower than 2.5G APD, likely attributable to thinner multiplication layer and higher electric field.

One method to improve the ESD is by adding ESD compensation in the module circuit [70]. The common examples of ESD compensation circuits include resistor and protection diode [72–74]. For APD, the ESD threshold can be improved by use of a compensation circuit where the transimpedance amplifier (TIA) circuit contains a resistor connected in series, as shown in Fig. 7.19. The ESD benefit of the resistor may be attributed to lessened voltage drop across the photodiode.

Table 7.5 shows the comparison of reverse HBM ESD threshold of the mesa-type 10G APD with and without the TIA circuit. By use of the TIA circuit, the reverse ESD threshold of mesa-type APD can be improved from −150 V to −400 V, by a factor of $\times 2.7$. With the additional serial resistor and parallel capacitor, the ESD threshold can be further enhanced to −800 V.

**Fig. 7.19** Schematics of ESD improvement at module-level by use of a compensation circuit for PD or APD. A TIA circuit with a resistor connected in series [70]**Table 7.5** Reverse HBM ESD threshold of mesa-type 10G APD with and without the TIA circuit

Device HBM	Mesa-type APD Without TIA circuit	Mesa-type APD With TIA circuit	Mesa-type APD With TIA circuit + additional resistor
Reverse ESD threshold	−150 V	−400 V	−800 V

7.5 Conclusions

In this chapter, we review the state-of-the-art APD photodetectors for various applications including datacenter, wireless, PON, and telecommunication. Modern high-speed 10G and 25G APDs typically employ mesa-type structures with device shrinking. The mesa structures and thickness shrinking of the charge and multiplication layers may induce issues such as high electric field, Joule heating, and current crowding. Among those issues, high electric field is likely to impose most reliability risk.

The reliability implications of high electric field can be manifested in the following aspects.

(1) Optical/electrical overload

The overload damage threshold of APD generally decreases with increasing electric field. For 25G APD and beyond, the ever-increasing electric field is expected to impose further challenge for overload robustness.

(2) Aging degradation

Temperature and current are the two main factors that determine device lifetime. High electric field may impact the APD device lifetime by means of higher junction temperature or higher current density. Knowledge of the activation energy and current exponent are hence important for quantitative study of relationship between high electric field and device lifetime.

(3) ESD

Reverse ESD threshold of APD also shows general reduction with increasing electric field. For example, the reverse ESD threshold of 10G APD is typically lower (about -150 V) than that of 2.5G APD (about -500 V) due to higher electric field. For high-speed APD, ESD compensation circuit may be needed to meet the stringent ESD requirement.

References

1. D.B. Judd, G. Wyszecki, *Color in Business, Science and Industry*. Wiley Series in Pure and Applied Optics, 3rd edn. (Wiley-Interscience, New York, 1975)
2. B.E.A.Saleh, and M.C. Teich, *Fundamentals of PHOTonics*, 2nd edn. (Wiley-Interscienc, New York, 2007)
3. H.J. Haugan, S. Elhamri, F. Szmulowicz, B. Ullrich, G.J. Brown, W.C. Mitchel, Study of residual background carriers in midinfrared InAs/GaSb superlattices for uncooled detector operation. *Appl. Phys. Lett.* **92**(7), 071102 (2008)
4. M. Fukuda, *Optical Semiconductor Devices*, Chapter 4 “Photodiodes” (Wiley, New York, 1999)
5. J.C. Campbell, Recent advances in avalanche photodiodes. *J. Lightwave Tech.* **34**(2), 278–285 (2016)
6. J.S. Huang, H.S. Chang, Y.H. Jan, H.S. Chen, C.J. Ni, E. Chou, Temperature dependence study of mesa-type InGaAs/InAlAs avalanche photodiode characteristics. *Adv. Optoelectron.* Article ID 2084621, 1–5 (2017)

7. Y.H. Chen, J.M. Wun, S.L. Wu, R.L. Chao, Jack J.S. Huang, Y.H. Jan, H.S. Chen, C.J. Ni, H.S. Chang, E. Chou, J.W. Shi, Top-illuminated In_{0.52}Al_{0.48}As-based avalanche photodiode with dual charge layers for high-speed and low dark current performances. *J. Sel. Top. Quantum Electron.* **24**(2), 3800208 (2018)
8. M. Nada, Y. Muramoto, H. Yokoyama, T. Ishibashi, S. Kodama, InAlAs APD with high multiplied responsivity-bandwidth product (MR-bandwidth product) of 168 A/W-GHz for 25 Gbit/s high-speed operations. *Electron. Lett.* **48**(7), 397–399 (2012)
9. M. Huang, P. Cai, Su Li, L. Wang, T-I Su, L. Zhao, W. Chen, C-Y Hong, D. Pan, Breakthrough of 25 Gb/s Germanium on Silicon Avalanche Photodiode. *Optical Fiber Communications Conference, Technical Digest, paper Tu2D.2* (OFC, Anaheim, CA, 2016)
10. E. Ishimura, E. Yagyu, M.N. Kaji, S. Ihara, K. Yoshiara, T. Aoyaji, Y. Tokuda, T. Ishikawa, Degradation mode analysis on highly reliable guardring-free planar InAlAs avalanche photodiode. *J. Lightwave Tech.* **25**(12), 3686–3693 (2007)
11. J.S. Huang, Y.H. Jan, *Environmental Engineering of Photonic and Electronic Reliabilities: from Technology and Energy Efficiency Perspectives* (Scholars' Press, Saarbrücken, Germany, 2017)
12. J.S. Huang, H.S. Chang, Y.H. Jan, H.S. Chen, C.J. Ni, E. Chou, S.K. Lee, J.-W. Shi, *Highly Reliable, Cost-Effective and Temperature-Stable Top-Illuminated Avalanche Photodiode (APD) for 100G Inter-Datacenter ER4-Lite Applications, PHOTOPTICS* (Funchal, Portugal, 2018), pp. 119–124
13. Nokia mobile anyhaul, Nokia White Paper, 2017
14. Bloomberg, <http://fortune.com/2018/02/13/winter-olympics-2018-5g-networks/>
15. J. Horwitz, <https://venturebeat.com/2018/01/31/pyeongchang-will-host-first-major-5g-video-demonstrations-for-olympics-viewers/>
16. <https://www.fierewireless.com/japan-lays-more-groundwork-for-a-5g-tokyo-olympics>
17. <https://www.techradar.com/news/intel-and-ntt-docomo-prepare-5g-plans-for-tokyo-2020-olympics>
18. C.F. Lam, *Passive Optical Networks: Principles and Practice* (San Diego, Elsevier, 2007)
19. G. Kramer, *Ethernet Passive Optical Networks* (McGraw-Hill Communications Engineering, 2005)
20. Y. Miyazaki, T. Yamatoya, K. Matsumoto, K. Kuramoto, K. Shibata, T. Aoyagi, T. Ishikawa, High-power ultralow-chirp 10-Gb/s electroabsorption modulator integrated laser with ultrashort photocarrier lifetime. *IEEE J. Quantum Electron.* **42**(4), 357–62 (2006)
21. Y. Cheng, J. Pan, Y. Wang, F. Zhou, B. Wang, L. Zhao, H. Zhu, W. Wang, 40-Gb/s low chirp electroabsorption modulator integrated with DFB laser. *IEEE Photon. Tech. Lett.* **21**(6), 356–358 (2009)
22. J.W. Raring, L.A. Johansson, E.J. Skogen, M.N. Sysak, H.N. Poulsen, S.P. DenBaars, L.A. Coldren, 40-Gb/s widely tunable low-drive-voltage electroabsorption-modulated transmitters. *J. Lightwave Tech.* **25**(1), 239–348 (2007)
23. Y. Miyazaki, H. Tada, S. Tokizaki, K. Takagi, Y. Hanamaki, T. Aoyagi, Y. Mitsui, +1 dBm average optical output power operation of small-chirp 40-Gbps electroabsorption modulator with tensile-strained asymmetric quantum-well absorption layer. *IEEE J. Quantum Electron.* **39**(8), 1009–1017 (2003)
24. G.M. Smith, K.A. McIntosh, J.P. Donnelly, J.E. Funk, L.J. Mahoney, S. Verghese, Reliable InP-based Geiger-mode avalanche photodiode array. *Proc. SPIE* **7320**, 1–10 (2009)
25. J.S. Huang, Y.H. Jan, H.S. Chen, H.S. Chang, C.J. Ni, E. Chou, Predictive reliability model of 10G/25G mesa-type avalanche photodiode degradation. *Appl. Phys. Res.* **8**(3), 66–74 (2016)
26. J.S. Huang, Y.H. Jan, H.S. Chang, C.J. Ni, E. Chou, S.K. Lee, H.S. Chen, J.W. Shi, Nanoscale III-V semiconductor photodetectors for high-speed optical communications, Chapter 3, in *Two-dimensional materials for photodetector*, ed. by P. K. Nayak. (Rijeka, Croatia, InTech Open, 2018), pp. 49–73. ISBN 978-953-51-3952-2
27. H.C. Neitzert, V. Cappa, R. Crovato, Influence of the device geometry and inhomogeneity on the electrostatic discharge sensitivity of InGaAs/InP avalanche photodiode, in *EOS/ESD Symposium* (Santa Clara, CA, 1997)

28. M. Nada, T. Yoshimatsu, Y. Muramoto, H. Yokoyama, H. Matsuzaki, Design and performance of high-speed avalanche photodiodes for 100-Gb/s systems and beyond. *IEEE/OSA J. Light-wave Technol.* **33**(5), 984–990 (2015)
29. C.L.F. Ma, M.J. Deen, L.E. Tarof, J. Yu, Modelling of breakdown voltage and its temperature dependence in SAGCM InP/InGaAs avalanche photodiodes, in *IEEE Electron Devices Meeting (IEDM, San Francisco, 1994)*, pp. 22.5.1–22.5.4
30. J.S. Huang, H.S. Chang, and Y.H. Jan, “Reliability challenges of nanoscale avalanche photodiodes”, *Open Access J. Photoenergy*, (2017), p. 0015
31. C.R. Viswanathan, *Physical Principles of Semiconductor Devices* (EE Class Note, UCLA, 1993), pp. 172–208
32. C. Hu, PN and metal-semiconductor junctions, in *UC, Berkeley, EE Class Note, Chapter 4* (2009), pp.89–156
33. T. Bendib, L. Pancheri, F. Dieffal, G.-F.D. Betta, Impact of temperature and doping concentration on avalanche photodiode characteristics. in *Proceed. World Congress Engineering*, vol. I (WCE, London, 2014), pp. 5–8
34. D.J. Massey, J.P.R. David, G.J. Rees, Temperature dependence of impact ionization in submicrometer silicon devices. *IEEE Tran. Electron Dev.* **53**, 2328–2334 (2006)
35. L.L.J. Tan, D.S.G. Ong, J.S. Ng, C.H. Tan, S.K. Jone, Y. Qian, J.P.R. David, Temperature dependence of avalanche breakdown in InP and InAlAs. *IEEE J. Quantum. Electron.* **46**(8), 1153–1157 (2010)
36. M.S. Tyagi, Zener and avalanche breakdown in silicon alloyed p-n junctions. *Solid State Electron.* **11**, 99–128 (1968)
37. J.S. Laird, T. Hirao, S. Onoda, H. Ohyama, T. Kamiya, Heavy-ion induced single-event transients in high-speed InP-InGaAs avalanche photodiodes. *IEEE Trans. Nuclear Sci.* **50**(6), 2225–2232 (2003)
38. A. Alpert, High-speed jitter testing of XFP transceivers, in *Viavi White Paper* (2015)
39. A.S. Oates, Reliability of silicon integrated circuits. Chapter 7 in *Reliability Characterisation of Electrical and Electronic Systems*, ed. by J. Swingler (Woodhead Publishing, Cambridge, 2015)
40. C.V. Thompson, J.R. Lloyd, Electromigration and IC interconnects. *MRS Bull.* 19–25 (1993)
41. J. Proost, K. Maex, L. Delaey, Electromigration-induced drift in damascene and plasma-etched Al(Cu). II. Mass transport mechanisms in bamboo interconnects. *J. Appl. Phys.* **87**, 99–109 (2000)
42. A.S. Oates, M.H. Lin, Electromigration failure distributions of Cu/low-k dual-damascene vias: impact of the critical current density and a new reliability extrapolation methodology. *IEEE Trans. Dev. Mater. Rel.* **9**(2), 244–254 (2009)
43. K.L. Lee, C.K. Hu, K.N. Tu, In-situ scanning electron microscope comparison studies on electromigration of Cu and Cu(Sn) alloys for advanced chip interconnects. *J. Appl. Phys.* **78**, 4428–4437 (1995)
44. Hu Chenming, *Modern Semiconductor Devices for Integrated Circuits* (Pearson Education, New York, 2009)
45. J.W. McPherson, Time dependent dielectric breakdown physics-models revisited. *Microelectron. Reliab.* **52**, 1753–1760 (2012)
46. K.N. Tu, Recent advances on electromigration in very-large-scale-integration of interconnects. *J. Appl. Phys.* **94**(9), 5451–5173 (2003)
47. M. Yunus, K. Srihari, J.M. Pitarresi, A. Primavera, Effect of voids on the reliability of BGA/CSP solder joints. *Microelectron. Reliab.* **43**, 2077–2086 (2003)
48. A. Topol, D.C.L. Tulipe, L. Shi, D. Frank, K. Bernstein, S. Steen, A. Kumar, G. Singco, A. Young, K. Guarini, Three-dimensional integrated circuits. *IBM J. Res. Develop.* **50**(4/5), 491–506 (2006)
49. Y.J. Chang, C.T. Ko, K.N. Chen, Electrical and reliability investigation of Cu TSVs with low-temperature Cu/Sn and BCB hybrid bond scheme. *IEEE Electron Dev. Lett.* **34**(1), 102–104 (2013)

50. Y. Ma, Y. Zhang, Y. Gu, X. Chen, Y. Shi, W. Ji, S. Xi, B. Du, X. Li, H. Tang, Y. Li, J. Fang, Impact of etching on the surface leakage generation in mesa-type InGaAs/InAlAs avalanche photodetectors. *Op. Express* **24**(7), 7823–7834 (2016)
51. H. Sudo, M. Suzuki, Surface degradation mechanism of InP/InGaAs APD's. *J. Lightwave Tech.* **6**(10), 1496–1501 (1988)
52. U.R. Bandi, M. Dasaka, P.K. Kumar, Design-in reliability for communication designs, in *43rd ACM/IEEE Design Automation Conference* (San Francisco, CA, 2006)
53. J.S. Huang, Design-in reliability of modern wavelength-division multiplex (WDM) distributed feedback (DFB) lasers. *Appl. Phys. Res.* **4**(2), 15–28 (2012)
54. P.S. Ho, T. Kwok, Electromigration in metals. *Rep. Prog. Phys.* **52**(3), 304–348 (1989)
55. J.S. Huang, T.L. Shofner, J. Zhao, Direct observation of void morphology in step-like electromigration resistance behavior and its correlation with critical current density. *J. Appl. Phys.* **89**(4), 2130–2133 (2001)
56. E. Ishimura, E. Yagyu, M. Nakaji, S. Ihara, K. Yoshiara, T. Aoyagi, Y. Tokuda, T. Ishikawa, Degradation mode analysis on highly reliable guardring-free planar InAlAs avalanche photodiodes. *J. Lightwave Tech.* **25**(12), 3686–3693 (2007)
57. A.A. Efremov, N.I. Bochkareva, R.I. Gorbunov, D.A. Lavrinovich, Y.T. Rebane D.V. Tarkhin, Y.G. Shreter, Effect of the joule heating on the quantum efficiency and choice of thermal conditions for high-power blue InGaN/GaN LEDs. *Phys. Semi. Device* **40**(5), 605–610 (2006)
58. J.S. Huang, Reliability of optoelectronics. Chapter 6 in *Reliability Characterisation of Electrical and Electronic Systems* (Cambridge, UK: Woodhead Publishing, 2015)
59. G.K. Wachutka, Rigorous thermodynamic treatment of heat generation and conduction in semiconductor device modeling. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **9**(11), 1141–1149 (1990)
60. K.N. Tu, Y. Liu, M. Li, Effect of Joule heating and current crowding on electromigration in mobile technology. *Appl. Phys. Rev.* **4**, 011101 (2017)
61. K.N. Tu, C.C. Yeh, C.Y. Liu, C. Chen, Effect of current crowding on vacancy diffusion and void formation in electromigration. *Appl. Phys. Lett.* **76**(8), 988–990 (2000)
62. E.C.C. Yeh, W.J. Choi, K.N. Tu, Current crowding-induced electromigration in flip chip solder joints. *Appl. Phys. Lett.* **80**(4), 580–282 (2001)
63. J.S. Huang, E.C.C. Yeh, Z.B. Zhang, K.N. Tu, The effect of contact resistance on current crowding and electromigration in ULSI interconnects. *Mater. Chem. Phys.* **77**(2), 377–383 (2002)
64. J.S. Huang, Y.H. Jan, H.S. Chang, J. Chang, R. Chang, G. Liu, D. Ren, E. Chou, ESD polarity effect study of monolithic, integrated DFB-EAM EML for 100/400G optical networks. *CLEO-PR* (Singapore, July31-Aug.4, 2017). Paper#1018
65. <https://www.semiconductors.org/>
66. M.A. Saleh, M.M. Hayat, O.H. Kwon, A.L. Holmes, J.C. Campbell, B.E.A. Saleh, M.C. Teich, Breakdown voltage in thin III-V avalanche photodiodes. *Appl. Phys. Lett.* **79**(24), 4037–4039 (2001)
67. J.R. Black, Electromigration failure modes in aluminum metallization for semiconductor devices. *Proc. IEEE* **57**(9), 1587–1594 (1969)
68. J.S. Huang, Temperature and current dependences of reliability degradation of buried heterostructure semiconductor lasers. *IEEE Trans. Device Mater. Rel.* **5**(1), 150–154 (2005)
69. Generic reliability assurance requirements for optoelectronic devices used in telecommunication equipment. *Telcordia, GR-468-CORE* (2004)
70. J.S. Huang, T. Olson, E. Isip, Human-body-model electrostatic discharge and electrical over-stress studies of buried heterostructure semiconductor lasers. *IEEE Trans. Device Mater. Rel.* **7**(4), 453–461 (2007)
71. J.S. Huang, H. Lu, Size effect on ESD threshold and breakdown behavior of InP buried heterostructure semiconductor lasers. *Open Appl. Phys. J.* **2**, 5–10 (2009)
72. S.J. Chang et al., Improved ESD protection by combining InGaN-GaN MQW LEDs with GaN Schottky diodes. *IEEE Electron Device Lett.* **24**(3), 129–131 (2003)

73. Y.K. Su, S.J. Chang, S.C. Wei, S.M. Chen, W.L. Li, ESD engineering of nitride-based LEDs. *IEEE Trans. Device Mater. Rel.* **5**(2), 277–281 (2005)
74. J. Weinlein, D. Sanchez, J. Salas, Electrostatic discharge (ESD) protection for a laser diode ignited actuator, in *Sandia Report* (2003), pp. 1–17

Chapter 8

Miniaturized Surface Plasmon Resonance Based Sensor Systems—Opportunities and Challenges



Peter Hausler, Carina Roth, Thomas Vitzthumecker and Rudolf Bierl

Abstract Surface Plasmon Resonance (SPR) is a well-known and established technology in bioanalysis and pharmaceutical sciences. Due to the expensive instrumentation and the need of trained people, it is mainly limited to applications in laboratories. However, there are some areas like environmental monitoring, chemical processing and civil infrastructure, which urgently need new sensor technologies. SPR has the potential to serve these fields. In order to be qualified for a use in these areas SPR has to overcome some hurdles. The instrumentation has to be robust, small in size and cheap. A device, which fits these needs, will be a micro-opto-electro-mechanical system (MOEMS) with integrated intelligent algorithms. In this book chapter, examples of miniaturized SPR devices are introduced, the limitations which have to be overcome as well as the possibilities for future applications are proposed. Due to the manifold advantages of this technology and the dropping prices for imaging sensors, Surface Plasmon Resonance imaging (SPRi) might become one of the leading technologies for SPR smart sensor systems.

8.1 SPR Principle, Current Applications and Future Evolution

Surface Plasmon Resonance (SPR) is a well-established and frequently used technology in bioanalysis and pharmaceutical sciences [1–3]. It enables a high sensitive, label-free, and real-time detection of binding events of (bio-)molecules to functionalized surfaces. It provides information about the specificity of biological interactions, the concentration of an analyte in a complex solution as well as the binding kinetics between molecules. Therefore, SPR-devices are very popular in drug discovery, fundamental biological studies, genetically engineering and environmental monitoring. The development of SPR-imaging devices is enabling the parallel screening

P. Hausler (✉) · C. Roth · T. Vitzthumecker · R. Bierl
Sensorik-ApplikationsZentrum, Ostbayerische Technische Hochschule Regensburg,
Franz-Mayer-Str. 1, Regensburg, Germany
e-mail: Peter.Hausler@OTH-Regensburg.de

© Springer Nature Switzerland AG 2019
P. Ribeiro and M. Raposo (eds.), *Optics, Photonics and Laser Technology 2018*, Springer Series in Optical Sciences 223,
https://doi.org/10.1007/978-3-030-30113-2_8

of multiple analytes at the same time and a high throughput rate [4, 5]. However, SPR-systems are still facing some drawbacks like high asset costs and the need for a controlled environment for such a refractometric measurement system to overcome the influences of temperature fluctuations. Many of the existing devices are designed to be used in a lab and therefore trained operators are required. As a consequence, SPR-systems are not ready to be used as smart sensors. A sensor should be a small device with low costs, which converts a physical value into an analogue or a digital electrical signal (Fig. 8.1). For many sensors calibration, processing and the transfer of the signal as well as the interpretation of the signal remains in the user’s responsibilities. While a smart sensor handles these tasks autonomously, the users receive an answer, which is already processed and interpreted. A smart sensor is an integrated, intelligent device with self-calibration, self-control, signal conditioning and a digital interface. A classical sensor gives its signal to higher control layer near to the sensor, while a smart sensor transfers its information to a control centre, which could be far away (Fig. 8.2).

SPR-devices are characterized by excellent sensitivity, which is enabling a wide range of applications from water analysis to medical point of care devices or even controlling of chemical processes. However, SPR-systems suffer from large size

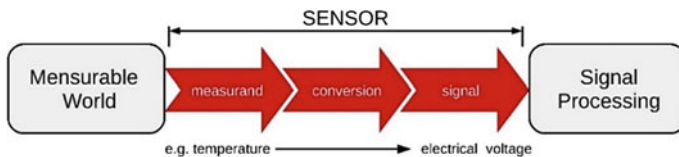


Fig. 8.1 Definition of a sensor. A sensor is converting a physical value into an analogue or a digital electrical signal. In order to receive a useful information, the signal output of the sensor has to be calibrated and interpreted

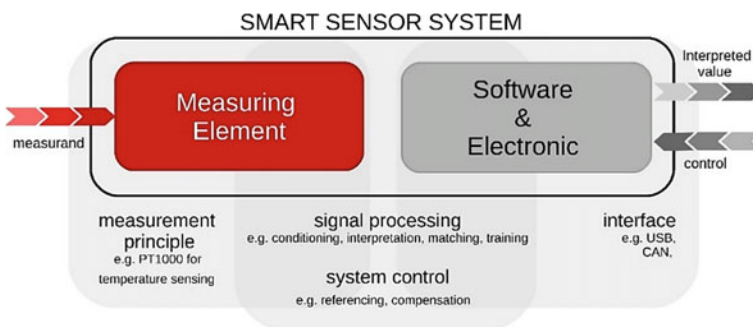


Fig. 8.2 Definition of a smart sensor system. A smart sensor system is self-referencing, it is able to compensate a drift, it is applying a signal conditioning and it is interpreting the signal from the sensing element. The system will provide a rated information after the signal processing. The interface to the user or next level of processing is digital, therefore the next instance does not have to be next to the sensor system, it could also be hundreds of kilometres away

and high costs. A simple downsizing by just scaling the systems would decrease the sensitivity without becoming cheaper. Presently there are significant efforts to develop new strategies for the miniaturization of SPR-devices. The next generation of SPR-devices will be a smart sensor system. To reach the next level several challenges, have to be overcome. The overall size of the system has to be significantly reduced. The device needs to compensate changes in temperature or avoid them. The same is valid for vibrations and other sources of interference. Therefore, robust referencing systems have yet to be developed.

If it comes to classical SPR-sensing, there exist four basic principles, which are angle modulation, intensity modulation, wavelength modulation and phase modulation. For a comparison of these techniques see [6]. All these techniques have individual advantages and disadvantages. For the use in compact smart sensor systems, intensity modulation is recommended since it is simple and robust in terms of instrumentation. There are no moving parts, which makes it less prone to vibrations and environmental stress. Some more technologies can be found in [7].

Prism based SPR-imaging sensors, which combine the intensity modulation with a motionless angle scanning technology are recommended. One can modify the sensor surface with different types of receptors and the size of the regions of interest (ROI) can be varied in a wide range. The typical size varies from $50 \mu\text{m}^2$ to 1cm^2 [8]. The freedom in arranging a big number of receptors and reference regions enables different referencing and calibration technologies [9].

Figure 8.3 shows the principle of SPR-sensing. While a thin gold film is irradiated by light, typically the entire light will be reflected. However, if the light is p-polarised and the angle of incidence is altered, one can see a narrow dip in the intensity of the reflected light. This dip is indicating that at this certain angle of incidence SPR-angle

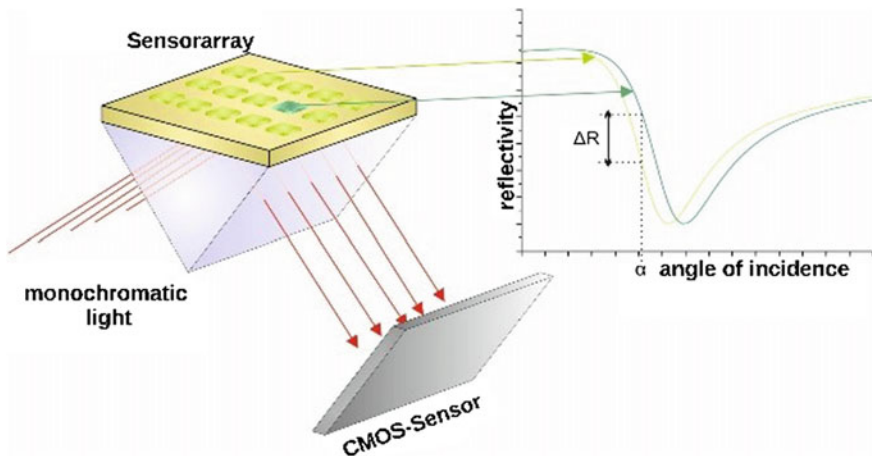


Fig. 8.3 Scheme of a SPR-sensor principle; the reflecting gold film consists of several regions of interests (ROI) in green; the diagram shows the signal response (ΔR) which is caused by a change of the refractive index in close proximity to each sensor spot [11]

surface plasmons are excited. The SPR-angle mainly depends on the refractive index in close proximity to the gold film, which is deposited on the coupler, usually a glass prism. Therefore, the refractive index on this side of the gold is constant, which means that any variations in the chemical composition in the analyte matrix floating over the receptor surface (and therefore in the refractive index) is determining the position of the SPR-angle. Selectivity to a special molecule of interest is generated by a chemical functionalization of the gold film by recognition elements [10]. Setup SPR-imaging sensor system is shown in Fig. 8.4.

There are already a couple of approaches for miniaturized SPR systems with good characteristics for special applications, but none of them is capable for continuous operation over months or even years, as a miniaturized smart sensor system should be. A very ambitious use case for SPR-smart sensors would be the monitoring of large power transformers. They are operated at temperatures between -20 and $+80$ °C in field and they are lacking some sensors, which are able to monitor the condition of the insulating liquid preventing an electrical breakdown. A SPR-imaging system, which is able to operate at a wide temperature range, is shown in [11]. However, the most critical challenges which need to be overcome for long-time use of SPR-systems are surface fouling of the receptor surface [12, 13], aging of components like camera or light source [14], temperature drift [15] and durability against vibrations. There are already individual solutions for most of these issues, however the combination of all of them in one device has not been reported so far.

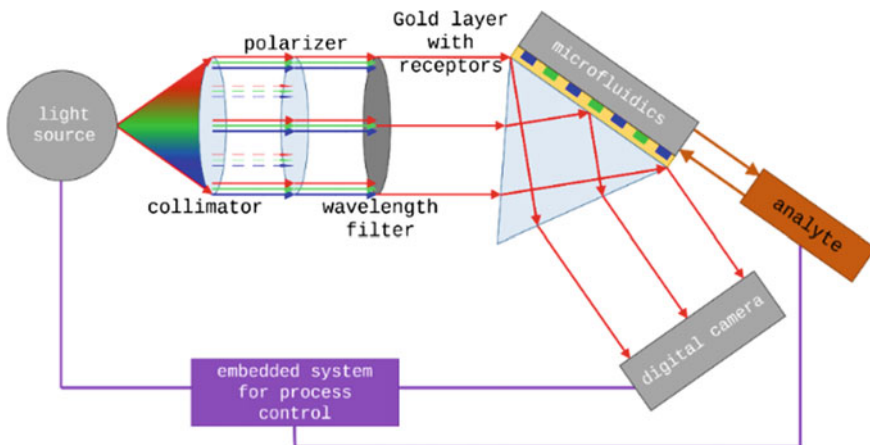


Fig. 8.4 System-setup, consisting of a light source, some beam shaping elements, a sensitive area with receptors as well as microfluidics, a camera detection system and a data processing unit

8.2 Current Approaches to Miniaturized SPR Systems

One can find reports on several miniaturized SPR-systems in literature and even on the market. However, these systems are not intended as smart sensors in a long-term scenario. Most of them are still intended to be used in lab in short time applications.

One of the miniaturized systems on the market is the Kmac SPRmicro system. It is a small system which is intended for lab-use, it provides no built-in algorithms for compensation of environmental influence or aging, does not make use of a fluidic system, which acquires the sample automatically, and it is leaving the interpretation of the data to the user.

A miniaturized SPR-system which uses another technique is the MICO SPR NANO from MiCoBioMed. It consists of a rotating mirror, which limits further miniaturization and might be a source of trouble if it comes to outside influences like vibrations. It needs trained lab staff and alike the Kmac system it does not provide a built-in algorithm.

Another miniaturized system, which is able to perform an angle scan, is the BIO-SUPLAR 6 from Analytical μ -Systems. It is rotating the SPR-prism in order to perform the angle scan. Therefore, it has a wideangle range of 17° . However, it is intended for lab use and it has moving parts, which makes a further miniaturization complicated.

One of the first commercially available miniaturized devices was the Spretta from Texas Instruments. The promising concept of this optical system was convincing by a highly integrated and very compact design. However, it also did not compensate environmental influences, has no automated sample acquisition and no built-in algorithms. Some scientists adapted the Spretta system by trying to compensate the lack of these issues [16]. Thereby the system became obviously bigger and only some of the drawbacks could be rectified (Fig. 8.5).

Fig. 8.5 SPR-system based on a texas instruments spretta with added temperature controller and peltier module for temperature regulation [16]



The SPR-systems on the market are measurement instruments for use in laboratories but they are no sensors, since the sample supply, the interpretation of the signal and the experimental procedure has to be done by lab staff. The main reasons for the absence of SPR smart sensor systems might be a lack of computing power and robust microfluidic systems in the past. However, some scientific systems already show different, promising ways to high-integrated smart sensor systems. All the approaches have advantages and disadvantages, which might be more or less important depending on the field of application.

The group of Homola reported in 2006 a novel surface plasmon coupler and disperser (SPRCD) technology, which is very promising from the perspective of miniaturization [17]. This approach does not need a prism and it is working with a broadband light source, which facilitates the miniaturization. They already showed a compact laboratory device with multi-channel measurement and temperature stabilization [18]. However, this device is still intended for laboratory use, controlled by scientists (Fig. 8.6).

Masson reported in 2017 a portable SPR-system, which is intended to be used in field under changing environmental conditions [19]. It is based on the commercially available P4SPR System from Affinite. This approach is catching up the need for a transportable and robust SPR-system. The system was tested in field and it is able to work in a wide temperature range. RDX was detected (ppb concentration) in water samples, which were taken from the environment. However, it is still not a very compact smart sensor system, since it is a “Lab-in-a-Jeep” system (Fig. 8.7).

Already in 2008, [20] demonstrated a hand-held SPR device which is able to work without an external computer or power source. It was a great step towards portable SPR-devices and it demonstrated the possibility of miniaturizing very compact SPR-systems. However, the system was intended for clinical use and it needs trained people for operation (Fig. 8.8). The samples have to be prepared and injected manually. The system consists of a single channel and it is not working with SPR-imaging.

Fig. 8.6 A laboratory prototype of a compact SPRCD sensor. The system has temperature control in order to suppress the effects of the ambient temperature change [18]

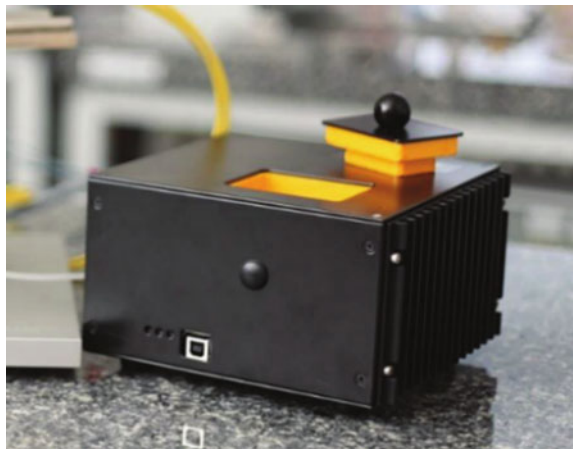




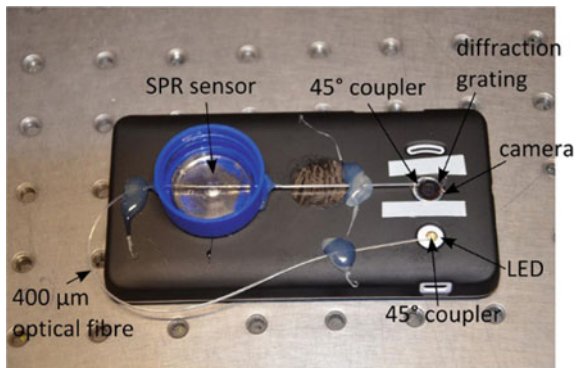
Fig. 8.7 Lab-in-a-Jeep SPR-system [19]

Fig. 8.8 A portable SPR-sensor, which can operate in hand-held mode. The sample loading occurs manually via the black lid port. The readout is via the LCD screen. The device has a built-in battery [20]



In 2015, Bremer reported the first smartphone-based fibre coupler SPR-sensor [21]. It is a very basic, compact and cost-effective approach (Fig. 8.9). The setup demonstrates the potential for low cost and disposable SPR-sensor. However, it has

Fig. 8.9 A fibre optic SPR-sensor system realised by [21]. The device shows how a SPR-sensor could be designed in a very basic way



no built-in referencing or calibration system and it is no autonomous smart sensor system.

Almost at the same time, Wei Peng reported a more advanced smartphone based SPR-sensor, which is using a fibre optics coupler [22]. It is a very compact, low-cost system, which demonstrates that it is possible to develop very small sensor systems. The approach is utilizing the smartphone's light source and camera for excitation and detection. The light from the smartphone's LED is coupled into an optical fibre after it passed a filter (Fig. 8.10). The cladding of the fibre is removed at a small part of the fibre and a gold film is coated onto that sensing region. It has a measurement channel, a control channel and a reference channel. Therefore, it should be able to compensate some of the environmental disturbances. However, it still is a system for trained users, short time usage and it has no automated sample acquisition and data interpretation.

The most advanced step towards a smart sensor system based on SPR-imaging was shown by [23]. They realized a compact SPR-imaging device based on a grating coupler (Fig. 8.11). The grating coupler was fabricated of a commercial Blu Ray® disc, which is very cost effective and might be a promising way for low cost fabrication. The imaging approach is more desirable compared to single-spot since it can be used for detection of a very big number of analytes in parallel. SPR-imaging is able to perform the same measurements several times on the same sensor chip together with smart algorithm ending up in higher reliability, precision and on-chip self-calibration. Therefore, this approach is a very promising route to a smart sensor system based on SPR-imaging.

However, the system is still not capable to work autonomous, for a long-term and without qualified personal. The sample acquisition has to be done fully manually. Because of the grating coupler, the light has to travel through the analyte before it reaches the metallic layer and finally the detector. This is giving rise to an additional source of interference as any absorbance or scattering of the light will cause

Fig. 8.10 Fibre based SPR-sensor installed on an Android smartphone. The camera of the smart phone captures images of the measurement as well as of the control to plot the data at the screen of the phone [22]



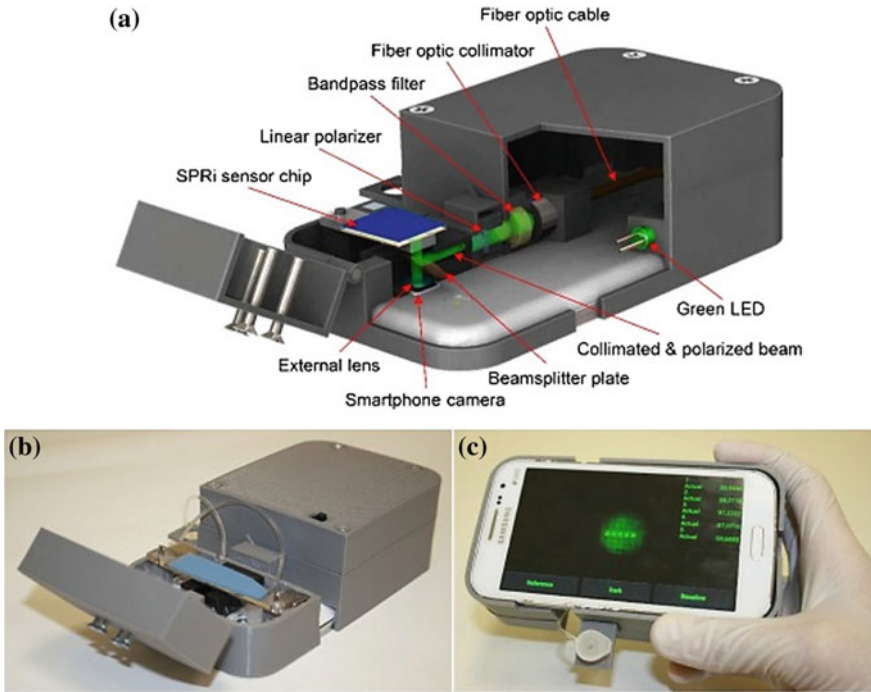


Fig. 8.11 Surface plasmon resonance imaging platform integrated with a smartphone. **a** Schematic illustration and **b** photograph of the imaging apparatus. **c** Smartphone App for real-time monitoring of multiple sensing spots. The SPR-system is based on a grating coupler [23]

fluctuations in the excitation power density of the light. Turbidity will strongly affect the signal. Furthermore [24] assert that prism coupler in general will result in higher sensitivity.

Smartphones are known for a very short product life cycle and they come in individual designs and sizes. Therefore, it is difficult to imagine a product suitable to be used for many years. Moreover, smartphones have many functionalities which are energy, power and money consuming and which are not necessary for a smart sensor system. For these reason, smartphones will not be the platform of choice for industrial applications. However, they can be an attractive alternative for the use at a home care scenario.

All miniaturized SPR-systems so far were designed for a narrow or very specific field of application. However, many industrial applications are lacking for sensors, which are capable of monitoring processes and environments. For the use in these highly complex matrices SPR smart sensor systems might change the gears and SPR-sensor technology will become even more popular.

8.3 Technical Aspects for the Design of Miniaturized SPR Sensors

One of the biggest problem in practical applications of the systems described so far is the need for the user to deal completely with the sample preparation and its application to the sensor. This is prone for generating not accurate results. Therefore, the use of a professional platform is mandatory and beneficial for smart sensor systems. For developing such a system, a x86 or an open ARM platform with Windows is recommended, since it is very flexible, and it is supporting higher programming languages like LabVIEW. However, for embedded systems the use of a Linux based operating system is advantageous, as it provides considerable higher security level. Open source is not only limited to software, there is also open hardware available, e.g. the Arduino platforms or the Raspberry PI eco systems. The Raspberry PI already offers a development kit for vision systems, which is supported by Google. This kit provides a coprocessor, which is optimized for neural networks. Open hardware and open software has left its infancy and turned from a toy for geeks to tools for professionals [25]. Transparency, which is a main feature by open source software like Linux will be a unique selling point if it comes to serve critical infrastructures like power networks, drinking water or military and even governments. Moreover, a transparent infrastructure is facilitating the implementation of new code and interfaces since it can be done by the developer itself and not only by the provider of the code. Some high-level programming languages like Python are also able to run on different operating systems like Linux, Windows and Mac.

For applications in biotechnology and medicine, SPR-devices have been equipped with a very limited number of receptors like DNA, proteins or polymer layers. The use of such systems is restricted to already known scenarios. At present stage diagnosis is mostly done by sensing markers in case of any health issue, by doing a single measurement when a person is ill. This is already outstanding, but smart sensor technology can bring SPR a step forward. From smart watches and fitness tracker, which got very popular in the last two years, one can clearly see the trend that people want to monitor their health constantly. For such purpose, a smart SPR sensor would be a promising tool to monitor changes in body fluids. One can imagine detecting even unknown markers which suddenly appear, or which concentrations are changed in body fluids to be able to detect any health disorder much earlier or give more accurate data to the doctor for better diagnosis. An SPR-imaging system with a sensing field of 10×10 mm and a spot size of $100 \mu\text{m}$ with a distance of $50 \mu\text{m}$ between the ROIs generates about 4,300 ROIs. Even if one third of the ROIs is needed as referencing spots and the other third are redundant ROIs there are still 1,000 ROIs left on one single chip. If one generates such a large variation of differently modified semi-selective receptors it is expected to screen and monitor a lot of marker molecules simultaneously or see if there are changes in the composition of fluids like blood, urine, saliva or tear fluid over time, if these changes are reversible and correlate these data with the health condition. The data analysis of such a system would be an endless work for any doctor. Therefore, anonymized databases of signal

patterns collected from many such sensors will be needed. They will be used to train neuronal networks in order to be able to indicate from a special pattern to a health risk. For the technical data analysis on a single device Deep Neural Networks (DNN) would have no problem to train the sensor under real conditions. In terms of image processing, especially Convolutional Neural Networks (CNN) are a very useful tool. Presently Tensorflow is the most used framework for deep learning. Google supports Tensorflow and it is open source, it was designed to be used with Python. An analysis by The Data Incubator shows clearly the dominance of Tensorflow in the field of deep learning frameworks [26].

Fouling and pollution is always an issue when it comes to long-time real-world application of a sensor system. For lab applications, this can be easily overcome or controlled, for continuous monitoring of real samples without sample pre-treatment this issue is almost impossible to be solved exclusively by referencing strategies. However, it can be recognized and subsequently eliminated by a cleaning system. Most of the actual SPR-systems do have a very simple microfluidic system, which is not able to perform complex operations. Since SPR is very surface sensitive it will react extremely to sediments and adhesions. Therefore, it will be necessary to include some kind of cleaning or regeneration mechanism.

Up to now, SPR-imaging systems have spatially separated components. They are arranged in a housing and are individually fixed to the housing or to a frame. If smart sensors based on SPR-imaging technology should be available to a broad spectrum of applications, these set-ups have to be replaced by a single MOEMS. Light source, camera, optical coupler and beam shaping elements have to be combined in one module for the reason of lowering the fabrication costs and increasing the robustness by additionally smaller size of the whole sensor system.

Most of the SPR-systems reported so far are optimized to be used with aqueous systems characterized by refractive indices smaller than $n = 1.4$. However, for applications in quality control of industrial processes, the refractive index of the matrix might be higher or chemically aggressive or capable of creeping, which needs to be kept in mind and which will need further improvements and changes to the systems as known in the present stage.

In the following, there will be a discussion of individual parts necessary to construct an SPR-imaging system and which parameters are important with respect to ending up with a smart sensor system.

(1) Characteristics of 2D Photon Detectors from a SPR Perspective

The detector is one of the most important parts of a SPR-sensor. Together with the light source, it is limiting the resolution of the device [27]. When choosing a camera, there are a few characteristics like full well capacity, temporal dark noise, dynamic range, quantum efficiency, absolute sensitivity threshold and signal-to-noise ratio which need to be taken into account. Another parameter which is also important but not directly connected to the camera's sensor chip is the interface. As all modern cameras are digital cameras, also the bit number of the A/D converter cannot be neglected.

Full Well Capacity. Each pixel of a camera can be imagined like a well holding electrons. The full well capacity shows the maximum number of electrons which one pixel can store. Larger pixels usually do have a larger full well capacity. A higher full well capacity enables a higher dynamic range. If a pixel reaches its full well capacity, it will no more change its value even if the light intensity increases. In practice the full well capacity is nearly the same as the saturation capacity even if there are some differences which can be found elsewhere [28].

Temporal Dark Noise. Random electrons are created which will be trapped by the cameras pixels (wells) where they were detected and turned into a signal. The dark noise electrons are generated by temperature. Temporal dark noise origins from the camera itself, which can be reduced by cooling the camera.

Photon-Shot-Noise. The shot noise is describing the fluctuation of the number of photons, which are detected by a pixel. Not every photon, which reaches a pixel, will be converted into an electron. The number of photons, which are converted into electrons, are subjected to a Poisson distribution. Therefore, the shot noise is also related to the quantum efficiency. While watching a source of light, the number of photons is not constant. The number of photons, which reach the pixel, is subjected to a Poisson distribution. The combination of these two effects is called shot noise. There are some more sources of noise, for a detailed explanation see [29].

Dynamic Range. The dynamic range of the camera is limited at the upper end of the range by saturation of the signal (saturation capacity) and at the lower end by dark noise (see Fig. 8.12). A camera with higher dynamic range can detect more

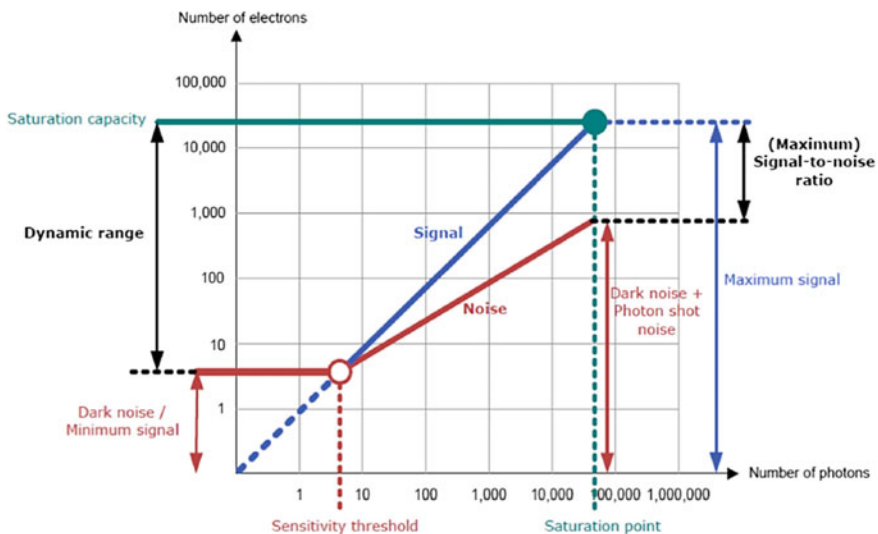


Fig. 8.12 Photon transfer curve of a digital camera. The number of photons has to exceed the dark noise in order to generate a usable signal. The dynamic range, which is often used to characterize a camera, is the quotient of saturation capacity and dark noise. From *Handbuch der Bauelemente der Optik* [29]

detail in the dark and bright condition. In situations, where both very bright and dark areas are being imaged at the same time, a higher dynamic range is beneficial as high dynamic range cameras are able to gather data in very bright and very dark areas. Camera systems with a big dynamic range are capable of recording images in dark environment as well as in very bright environment. A high dynamic range is also very useful if one is doing classic angle scan SPR. It enables a very fine resolution of the SPR-slope.

Absolute Sensitivity Threshold. The absolute sensitivity threshold is the minimum number of photons, which is needed to equal the cameras noise. The absolute sensitivity threshold equals the dark noise in Fig. 8.12. A signal, which is lower than this value, cannot be detected by the camera.

Signal to Noise Ratio. The signal-to-noise-ratio (SNR) is comparing the level of the signal used to the level of the background noise. The value of the SNR is the fraction of the whole signal over the sum of all noise.

Quantum Efficiency. The quantum efficiency is the ability of the camera to turn photons into electrons. A higher quantum efficiency means greater sensitivity for incoming light. A quantum efficiency of 80% means that for every 100 photons that hit the camera an average of 80 will be detected and turned into a signal. The quantum efficiency depends on the wavelength of the incoming light. Figure 8.13 shows a comparison of the quantum efficiency of two typical CMOS camera chips. It is obvious to see that the quantum efficiency at 500 nm is twice as high as at 700 nm. Therefore, one should always take into account the quantum efficiency at the desired wavelength and not only the one around 500 nm that is often used for EMVA 1288 data sheets. The EMVA 1288 is a standard designed to make the performance of cameras comparable.

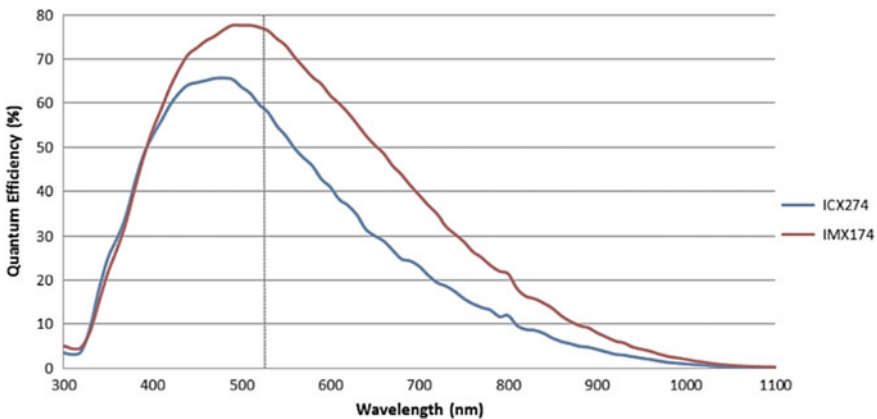


Fig. 8.13 The quantum efficiency is a measure of the sensor's ability to convert photons to electrons and depends on the wavelength, which is used to enlighten the sensor. From FLIR Integrated Imaging Solutions, Inc. (<https://www.ptgrey.com/KB/10795>)

Camera Interface. There are many different camera interfaces on the market, which all have advantages and disadvantages. For small smart sensor devices like an integrated SPR-sensor an interface that requires a large capture board will not be suitable. Nevertheless, one can take advantage of a high data transfer rate and a small and common interface as e.g. known from the USB3 interface. USB3 interfaces are even in the smallest stick PCs available, which makes them attractive for integrated devices. Moreover, USB3 Vision cameras can easily be integrated into LabVIEW software because of the NI-IMAQdx driver and the GenICam standard a general interface. It is also very easy to switch between different USB3 Vision cameras from different manufacturers since they all do support the same software interface. A comparison of the most common camera interfaces is shown at Fig. 8.14.

Number of Bits. The number of bits is limiting the resolution of the SPR-system. For example, a signal change from the maximum to the minimum of the dynamic range at a 10-bit system discriminates in 1,024 steps. By using a 14-bit system, one can already discriminate 16,384 steps. If a refractive index unit (RIU) change of 10^{-2} would cause a light intensity change which covers the whole dynamic range of a 10-bit camera, the smallest step which can be detected would be 10^{-5} . If a refractive index unit (RIU) change of 10^{-2} would cause a light intensity change which covers the whole dynamic range of a 14-bit camera, the smallest step which can be detected would be $6 \cdot 10^{-7}$. In practice, the observed change in RIU will never cover the full dynamic range of the camera. Figure 8.15 visualizes the influence of the number of bits on the resolution of the SPR-system.

(2) **Characteristics of Light Sources from a SPR Perspective**






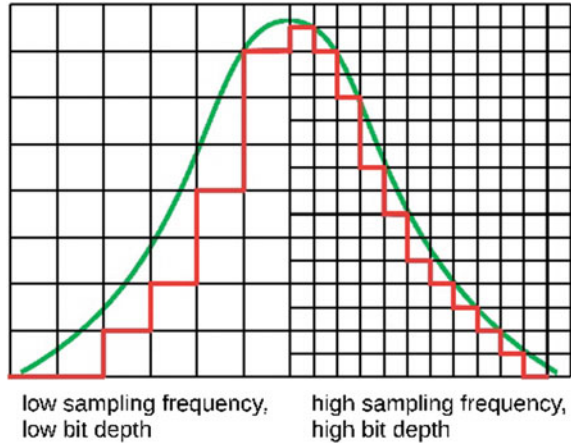
Digital Signal Options	FireWire 1394.b	Camera Link®	USB 2.0	USB 3.0	GigE
					
Data Transfer Rate:	800 Mb/s	3.6 Gb/s (full configuration)	480 Mb/s	5Gb/s	1000 Mb/s
Max Cable Length:	100m (with GOF cable)	10m	5m	3m (recommended)	100m
# Devices:	up to 63	1	up to 127	up to 127	Unlimited
Connector:	9pin-9pin	26pin	USB	USB	RJ45/Cat5e or 6
Capture Board:	Optional	Required	Optional	Optional	Not Required
Power:	Optional	Required	Optional	Optional	Required (Optional with PoE)

Fig. 8.14 Comparison of the most popular camera interface standards. Because of its high data transfer rate and its USB3 Vision software interface, the USB3 interface might be the best for developing a smart sensor system. Courtesy of Edmund Optics. All Right Reserved

Fig. 8.15 Influence of the sampling frequency and bit depth on the sensitivity of a SPR-system. With a rising bit depth, the smallest step, which could be resolved, is shrinking



For light sources the minimization of the noise and the selection of the wavelength are the main characteristics, but there are a lot of other parameters which have to be optimized when it comes to SPR-imaging sensors. Important parameters are the full width at half maximum, the polarization, the coherence, the homogeneity and the beam profile as well as the divergency.

Divergency. Well collimated laser light sources has nearly no divergency. This is giving rise to the possibility to acquire an image without a lens in front of the camera. According to Fig. 8.4 there has to be no lens between the prism and the CMOS/CCD chip to record an Image of the gold surface. The parallel light beam shines from the light source onto the gold surface where the ROIs cut out some of the intensity before the light is reflected. Finally, the parallel light beam reaches the camera without passing a lens, where it creates an image. The better the collimation of the laser the better the quality of the image. This technique should be a big advantage for miniaturized systems such as sensors. It is facilitating the design of MOEMS. Therefore, it is desirable to use a light source with no divergency.

Polarization. Only p-polarised light can create surface plasmons at the gold layer. If there is also s-polarised light it will pass the gold mirror without any change. This light will create an offset at the imaging chip, which in turn will lower the sensitivity of our system. Therefore, it is desirable to use high-polarised light source.

Coherence. When talking about coherence, we have to discriminate the different types of coherence, which are temporal coherence (Fig. 8.16) and spatial coherence [30]. Temporal coherence is a measure on how monochromatic a light source is. It tells us how well a wave can interfere with itself at a different time. The delay over which phase or amplitude wanders by a significant amount is defined as coherence time. The coherence time is large for monochromatic light and low for broadband light. If a broadband light starts the coherence is perfect but by time a shift occurs, the time until the shift will get significant is called coherence time, the distance which the light travels until the significant shift occurs is called coherence length.

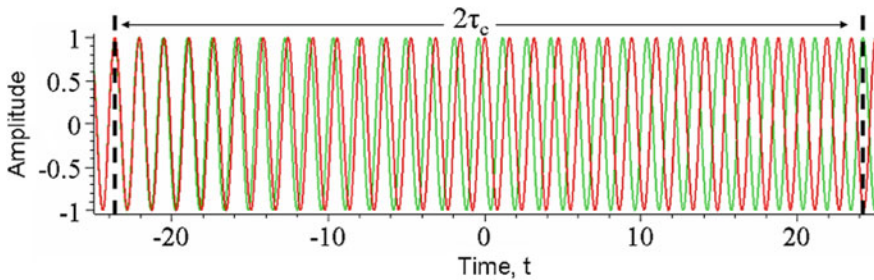


Fig. 8.16 One can see the amplitudes of two waves, whose phases (green & red) drifts significantly to each other by time (τ_c). The coherence time τ_c is the delay over which the phase or amplitude wanders by a significant amount of time. Graphic from J. S. Lundeen at Wikimedia Commons

Spatial coherence is a measure on how uniform the phase of a wave front is. The Youngs double slit experiment shows the spatial coherence. If the two holes are coincident or very closely spaced, the recombined waveforms at the detector are correlated so that they produce interference fringes. However, as the distance increases, the degree of correlation decreases and so does the contrast in the interference pattern. One of the phenomena, which can occur due to coherence, are speckles. To understand how speckles are generated one should imagine a flat surface, which has a fine roughness which is at the magnitude of the wavelength. Each point of the rough surface is acting as a source of a spherical wave. These waves are interfering at the far field. Another one is the diffraction, which occurs for example at dust particles, which get into the system. The diffraction of laser light is also used to determine the size of particle at specially designed setups [31]. However, it can be very disturbing for SPR-imaging systems.

Wavelength and full width at half maximum. The wavelength has a great impact on the SPR-effect. The refractive index of a material changes with the wavelength of the light [32]. Therefore, the whole system consisting of glass, metal layer and analyte must be matched carefully. Figure 8.17 shows how the refractive Index of Schott F2 glass changes with the wavelength.

Figure 8.18 shows the simulated SPR-slopes for a system consisting of Schott F2 glass, a 4 nm chromium adhesive layer, a 46 nm gold layer and water at different wavelengths.

It is obvious to see that the slope at 800 nm gives us the best result while the slope at 500 nm is the worst one. However, by choosing the best wavelength one should not forget about the quantum efficiency of the detector, which is often going down with increasing wavelength.

While talking about broadband lightsources like LEDs one also have to take into account the Full Width at Half Maximum (FWHM). It is a measure for the range of wavelengths a light source is emitting. Figure 8.19 shows how the FWHM is determined. It is the spectral width of a light source at the half of its maximum intensity. LEDs got a typical FWHM between 15 nm and 70 nm. If we take into account the big differences between the 600 nm slope and 700 nm slope at Fig. 8.18

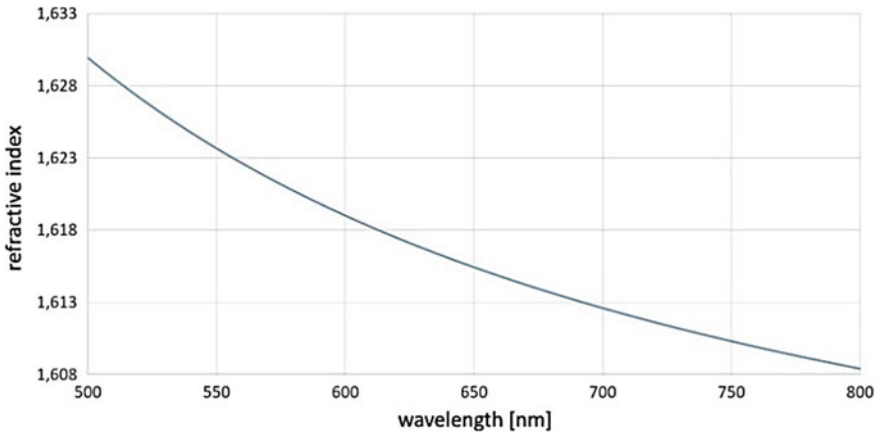


Fig. 8.17 Wavelength dependency of the Schott F2 glass. The refractive index of the glass is decreasing with an increasing wavelength

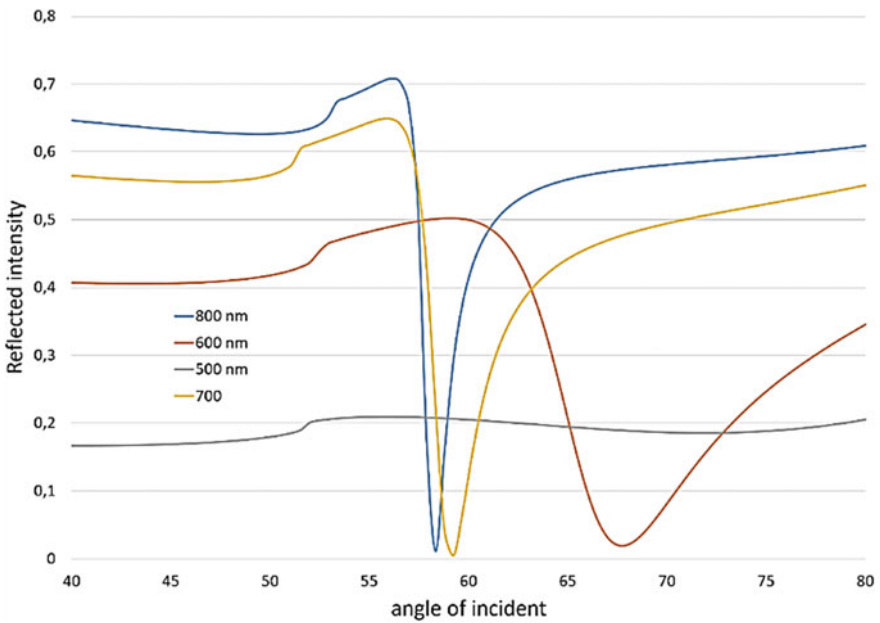
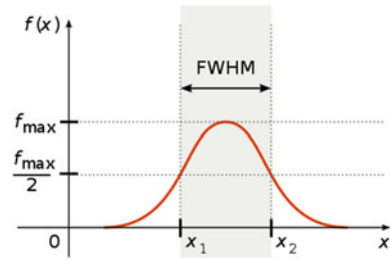


Fig. 8.18 Simulated SPR-curve for Schott F2 glass coated with a 4 nm chromium adhesive layer and a 46 nm gold layer. The analyte is water. Simulation was done with WinSpall

Fig. 8.19 Full Width at Half Maximum (FWHM). The FWHM is the width of the spectrum at the half of the amplitude. From Arne Nordmann from Wiki Commons (Creative Commons Attribution-Share Alike 3.0)



it is obvious that a FWHM of 50 nm will dramatically decrease our sensitivity. Therefore it is very desirable to have a very narrow FWHM.

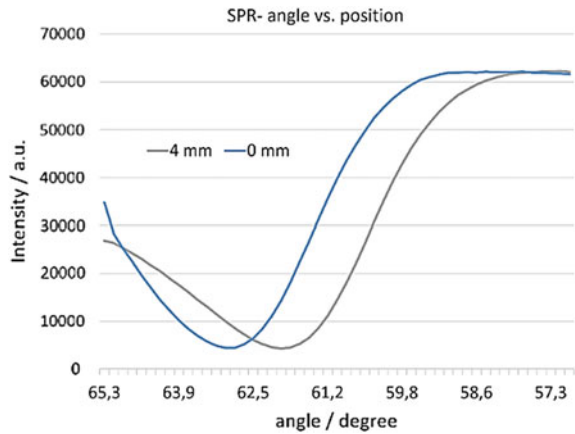
A narrow FWHM can be achieved by using a narrow band light source like a laser or by applying a narrow bandpass filter to a broadband lightsource.

Beam Shape and Divergency. Laser diodes have a gaussian beam profile. LEDs do have a similar profile which is depending on the structure of the LED. This implies that the gold surface is not enlightened evenly if only cylindrical lenses are used for beamshaping. A gaussian beam profile on our detector leads to a maximum intensity at the center of the beam profile and to a very bad intensity at the edges of the beamprofile. Therefore it is worthwhile to have a top hat beamprofile which can be achieved by different homogenization technologies. Depending on the homogenization technology one have to take into account that polarization can be lost. While homogenizing a laser lightsource one should also consider that collimation can be lost which would be a hard drawback if it comes to MOEMS which are working without camera lenses. If a camera lense is used divergency does not have to be a drawback. It can be used as an advance. If a divergent light source is used, the divergency will change the angle of incidence according to the propagation direction of the light. In other words there are lines with the same angle of incidence on the gold surface. Depending on the value of the divergence the angle is changing more or less.

Figure 8.20 shows two SPR-curves which were recorded at the same system, at the same time on the same gold surface at two different ROI's which were separated by 4 mm in direction of the light of incident. The shift occurs because of the divergency at the light of incidence. This technology gives rise to some very usefull methodes like motionless angle scan. On the other hand if somebody is using different receptors with different refractive indexes, every receptor can be placed on its optimum position at the gold surface according to its refractive index.

By considering all these requirements neither laser diodes nor LEDs are the optimum light source for all kind of setups. Depending on the most important requirements one have to accept some drawbacks. However, there is a third light source which should be taken into account since it is combining some properties of LED and laser. Superluminescent diodes have a narrower FWHM then LEDs, they can be collimated like a laser and they don't have coherence problems like a laser diode. The drawbacks of superluminescence diodes are a very high price compared to LEDs

Fig. 8.20 Two SPR-curves from the same gold surface, which were recorded at the same time in the same setup, are shown. The divergency of the light source, which was used, is spreading the angle of incidence over the gold surface. The two ROI's were separated by 4 mm [11]

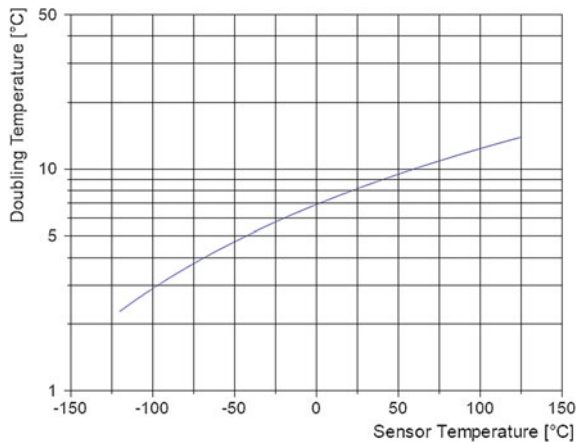


and a less pronounced polarization as well as a broader FWHM compared to laser diodes.

(3) The Influence of Temperature

The temperature of the system and the temperature stabilization of the system or the compensation of the fluctuation of the temperature is a very important part of a SPR-smart-sensor-system. The temperature is influencing the light source [33], the refractive index of the glass, the analyte [34] and the efficiency of the camera. Figure 8.21 shows the camera temperature versus the doubling temperature of the dark noise. The doubling temperature is indicating the temperature change, which doubles the noise of a camera sensor. If a sensor is cooled down to $-100\text{ }^{\circ}\text{C}$ a temperature change of $2\text{ }^{\circ}\text{C}$ will double the dark noise. If the camera is operating at $25\text{ }^{\circ}\text{C}$ a temperature change of $8\text{ }^{\circ}\text{C}$ is necessary to double the dark noise [35]. This

Fig. 8.21 Sensor temperature versus the doubling temperature of dark noise. The doubling temperature is indicating the temperature change, which is necessary to double the dark noise of a camera sensor



fact is very important for systems, which are operating in environmental conditions and should not be neglected.

(4) Microfluidics

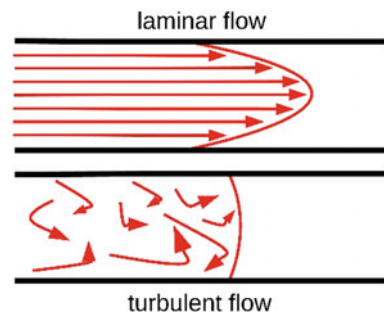
SPR-sensor systems can detect changes in refractive index next to the gold surface with an exponential decay in sensitivity up to distances of approximately half of the wavelength of the excitation light, e.g. about 320 nm for common 640 nm light source. It is desirable to build the fluidic system as small as possible to reduce the constructed space and fabrication costs and to provide a homogeneous flow of the analyte in this narrow volume next to the gold. It is of great importance not to trap air bubbles within this sensitive range, which will cause massive interference. There are many other reasons for small fluidics like a reduced consumption of analytes or a reduced impact on the analysed system. A reduction of the dimensions of the fluidic channel is accompanied with a laminar instead of a turbulent flow in nearly all scenarios (Fig. 8.22). Laminar flow induces a slower sample transport near to the borders of the channel, which could be undesirable. Depending on the materials used and liquids, also surface wetting needs to be taken into account.

The Reynolds number is a measure whether laminar or turbulent flow will occur.

$$Re = \frac{vd\rho}{\eta}$$

where ρ is the density and η is the viscosity of the fluid, d represents the diameter of the capillary and v is the velocity. The value where a laminar flow turns into a turbulent flow is called critical Reynolds number and is influenced by the geometry of the channel. For a straight channel, it is around 2,000. For a system where only the diameter of the channel is varied, one can see that the Reynolds number shows a linear change as a function on the channel diameter. Reducing the diameter of a channel from 1 mm to 100 μm , the Reynolds number would be reduced by a factor of 10. If the system would have a Reynolds number of 5,000 at 1 mm diameter, it would have a Reynolds number of 500 at a diameter of 100 μm . Therefore, the flow regime would turn from turbulent to laminar. Compared to the centre of the channel,

Fig. 8.22 Laminar flow versus turbulent flow. Most microfluidic devices will exhibit laminar flow since the flowrate is too low for turbulent flow. If it comes to laminar flow, the velocity of the fluid is lower at the surface of the channel



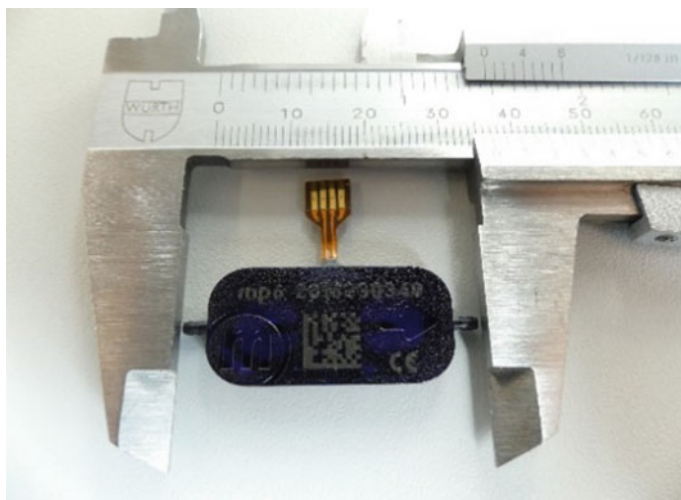


Fig. 8.23 Membrane micro pump. Membrane micro pumps are very small and space saving compared to other technologies. The displayed model is made of plastic but there are also some made of silicon and stainless steel

the flow rate of such a system is much lower at the border of the channel. Which in turn implies that the exchange of the analyte at the region with the highest sensitivity is slower.

Consequently, there will also be no mixture of different liquids, which are induced into a thin channel at different positions. If a mixing of the liquids is desired, one has to introduce a mixing device.

Hydrophobic or hydrophilic liquids show different behaviour in wetting towards materials, which are used for the fluidic system. Air bubbles can also be a problem, when get stuck on the gold surface where the channels usually get broader and the liquid will flow around the bubble. Therefore, bubble traps or filter systems for small particles should be integrated into a microfluidic system if samples are acquired from environmental conditions without any pre-treatment. The integration of such systems, which are able to work for a long time without satiation, is still challenging [36].

Liquids can be accelerated and transported in many different ways and there is a wide range of pumping mechanisms. Peristaltic pumps are very popular for lab use, since they do not interact with the analytes. However, the miniaturization of these devices is difficult and peristaltic pumps with higher flow rates are still very big. Membrane pumps can deal as alternative. They are cheap in purchase easy in use and small in size (Fig. 8.23).

However, microfluidics can get extremely complex in design and handling as fluids behave very different in small dimensions, but it has also a great potential in terms of miniaturized sensors. A detailed overview is given by Yujun Song [37].

(5) Data Processing and Compensation Strategies

The ROIs for data acquisition on the gold surface of the SPR-system can be very small. 50 μm are a typical size for small ROIs. While dealing with such small dimensions it is ambitious to align a system perfectly. Therefore, every individual ROI has to be matched to the camera image after system assembly, which can be a laborious work. Algorithms, which align the pattern automatically, need to be developed. In most cases, the ROIs will have geometrical shapes like a circle or a rectangle, which are very easy to detect [38]. However, higher programming languages like LabVIEW have got some VISION tools which are easy to use and are able to recognize individual shapes [39]. Higher programming languages are easy in use, which enables an even greater number of users to participate in software development. It allows everybody to integrate new components into the system with only a little expenditure of time. Standardization of interfaces should be considered for software interfaces as well as for hardware interfaces.

Temperature changes, fluctuations in the emission intensity of the light source at a certain wavelength due to limited lifetimes or aging, fouling and degeneration of the chemically modified gold surface with receptor- and reference-ROIs is challenging for SPR-imaging systems with the aim of long-term monitoring. These effects can be minimized or completely ruled out by smart compensation algorithms. A simple differential measurement between a reference and a receptor already partly compensate a drift of light source and camera or a temperature drift but it would not compensate an aging of the chemical receptor or reference. Therefore, it is necessary to develop algorithms, which are able to bias the system for these effects. Moreover, design strategies for complex microfluidics, which are able to recover the sensitivity by regeneration and calibration steps, are necessary. As one example, it is possible to imagine proving camera and light source by changing the polarization of the excitation beam from p- to s-polarization to check the intensity without generating any SPR-effect. Light pulses can also be used to desorb molecules by photoreactions from the surface, also to apply an electrical potential to the surface might be an option beside the development of robust receptor layers, which warrant a long-term stability.

However, the fields of image processing, neuronal networks and compensation strategies for small sensor devices are new to most SPR-researchers. Therefore, it would facilitate the development if there would be more research and standardization in this particular area.

(6) Choice of Material and Working Angle

Another important parameter is the choice of material for the prism as well as for the metal layer. The refractive index of the glass, which is used for the SPR-prism is related to the analyte matrix that is desired to be monitored. Highest sensitivity can be reached if the difference in refractive both, prism and matrix, is huge. By simulation, one can obtain a SPR-curve for the desired system. The refractive index of the matrix of interest usually cannot be changed, therefore the refractive index of the glass has to be chosen carefully to obtain the best sensitivity [40, 41].

Figure 8.24 shows a wide variety of glasses with very different refractive indexes, which are available to design an optimized SPR-system. Beside the refractive index, the chemical and physical capability of resistance should also be considered. While designing a multi wavelength SPR [42] system one should also mind the Abbe number, since it is determining how the refractive index is changing by a change of the wavelength.

Beside the type of glass, one should also mind the type of SPR-active-layer, since gold is the most prominent but not the only material, which is available [43]. A compromise between durability and sensitivity are gold-silver bilayers [44]. There are also differences between different layer deposition methods [45] and different types of adhesive layers [46]. Another possibility is to use bimetallic films. These have been described to lead to better sensitivity in affinity-based SPR-sensors [47, 48]. While using a silver-gold stack, the temporal stability of such thin layers in the nanometer regime has to be considered. An intermixing of the different ions with time will affect the SPR-signal in terms of long-time monitoring.

By structuring the metallic surface itself, it was shown that so-called nanohole arrays lead to a reduction of the plasmonic field and therefore the penetration depth of the sensing layer [49, 50]. It was shown that by such an approach it is possible to enhance the sensitivity in the detection of small molecules by SPR-technique [51, 52]. A comparison of SPR and LSPR can be found in [53].

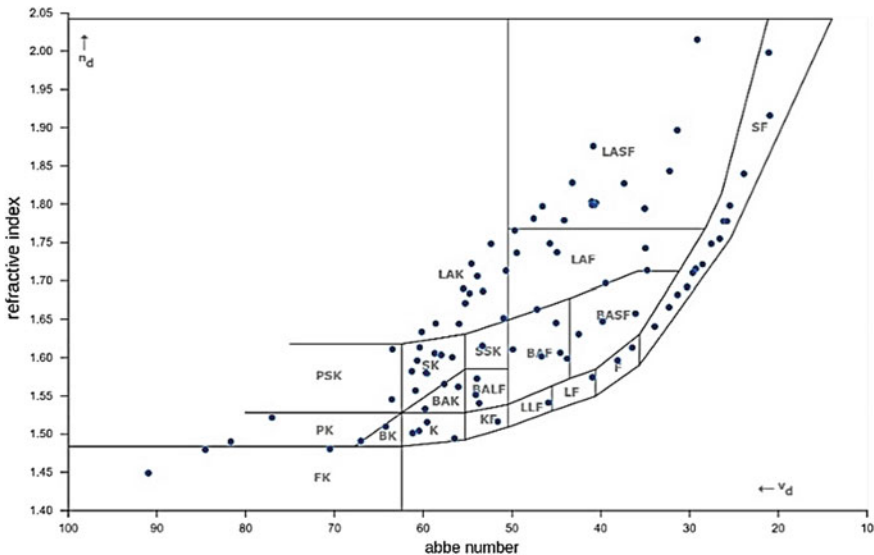


Fig. 8.24 A chart of different glass types available by Schott. It shows the refractive index versus the Abbe number. The Abbe number is a measure on how much the refractive index changes with the wavelength. The Abbe number is decreasing with a rising dispersion. From Schott AG

8.4 Outlook

The chemical industry is facing an upcoming change for a wide range of production sites. The trend is going from batch processing to continuous processing. In many cases continuous processing offers a better quality, more throughput, less waste, a better inventory management and less energy consumption [54]. For optimizing all these processes there is an emerging need for online and inline monitoring with a label-free technique. SPR-smart-sensor-systems are promising to bring a sensing technology for many of these scenarios where changes in the chemical composition of complex matrices needs to be monitored.

Another possibility for the need of such sensors are safety reasons for maintaining a high quality of the public water supply. This is especially very challenging in big cities or in countries with warm climate. Most of the time the origin of contaminations cannot be identified. If the water network would be equipped with sensors at all the intersections, it would be possible to notice and localize a contamination and its source very quick. This scenario, which origins from a natural source of contamination could also be extended to an attack on the water network. At the moment the analysis of drinking water takes a lot of time, the interval between the testing is usually very long and the costs are high. In many regions of the world, ground water is the basis for most of the drinking water. However, there are many sources of pollution, which are contaminating the ground water. Agriculture is deploying pesticides and fertilizers. Wastewater is containing drugs, hormones and radio-opaque substances and industrial wastewater can be even more toxic. In 2009 by analysing blood samples for blood donation it was found that blood from people in a certain region in Bavaria, Germany, contains an up to 20-times increased concentration of the cancerous perfluorooctanoic acid (PFOA). Much later it was found out that this chemical was used up to 2008 in industry next to this region, where it contaminated the sewage water and therefore over time the drinking water, which is still the case up to now. In 2018 the government needs to support breast-feeding mothers by analysing their milk for this toxic compound. By taking into account such long periods between cause and effect, it gets clear that a possibility to monitor the ground water for any changes in its composition over time will lead to better life as possible risks can be identified immediately [55].

The next evolution of SPR will arise due to the availability of tremendous computing power at small-scale systems and the rising megatrend of digitalization. Smart sensor systems are the senses of the digital world, they are distributed everywhere in the world and they are connected to each other. Smart sensor systems based on SPR will be a part of this sensing network. On the way to this goal, many hurdles have to be overcome. The system costs and the system size must be significantly reduced and new technologies like neuronal networks have to be integrated into the SPR-smart-sensor-systems. Such sensors will not only be equipped with a limited number of specific receptors for a small number of analytes. They can consist of up to 1,000 semi-selective receptors, which result in signal patterns, which will be analysed by neural networks. A constant machine learning process will even allow

detecting changes in a matrix which cannot be foreseen, and which will not be tested in a standardized lab analysis. Currently the most prominent smart sensor system, which is utilizing a deep learning network to discover information, which are not visible at first sight, is the apple smart watch, which is measuring the heart rate. By measuring the heart rate, it is able to detect atrial fibrillation, hypertension and sleep apnea.

The simplest SPR system, consisting of light source, sensing area and light detector only is prone to constant signal changes generated by the SPR-components and the environmental conditions. Many reasons are responsible for such signal drifts. At the moment every SPR-system developer and every SPR-user is working on compensation and avoidance strategies. Highly automated systems will offer a formidable number of compensation and corrective actions, e.g. use of many different references, redundant references and receptors, arranging references in different patterns, use of different light sources or wavelengths, or temperature compensation [9, 41, 42]. The number of possibilities is almost endless. It is very time consuming and it is holding a tremendous number of sources of error. Therefore, the outcome of the different strategies should be investigated and evaluated for their use for different problems. A review about how to stabilize a SPR-signal would definitely facilitate the work of the whole SPR-community.

References

1. J.F. Masson, Surface plasmon resonance clinical biosensors for medical diagnostics. *ACS Sens.* **2**(1), 16–30 (2017)
2. P. Singh, SPR biosensors: historical perspectives and current challenges. *Sens. Actuators B: Chem.* **229**, 110–130 (2016)
3. A. Olaru, C. Bala, N. Jaffrezic-Renault, H.Y. Aboul-Enein, Surface plasmon resonance (SPR) biosensors in pharmaceutical analysis. *Crit. Rev. Anal. Chem.* **45**(2), 97–105 (2015)
4. C. Liu, F. Hu, W. Yang, J. Xu, Y. Chen, A critical review of advances in surface plasmon resonance imaging sensitivity. *TrAC Trends Anal. Chem.* (2017)
5. R.B. Schasfoort (ed.), *Handbook of Surface Plasmon Resonance*. Royal Society of Chemistry (2017)
6. S.G. Nelson, K.S. Johnston, S.S. Yee, High sensitivity surface plasmon resonance sensor based on phase detection. *Sens. Actuators B: Chem.* **35**(1–3), 187–191 (1996)
7. G.A. Lopez, M.C. Estevez, M. Soler, L.M. Lechuga, Recent advances in nanoplasmonic biosensors: Applications and lab-on-a-chip integration. *Nanophotonics* **6**(1), 123–136 (2017)
8. M. Puiu, C. Bala, SPR and SPR imaging: Recent trends in developing nanodevices for detection and real-time monitoring of biomolecular events. *Sensors* **16**(6), 870 (2016)
9. D. Boecker, A. Zybin, K. Niemax, C. Grunwald, V.M. Mirsky, Noise reduction by multiple referencing in surface plasmon resonance imaging. *Rev. Sci. Instrum.* **79**(2), 023110 (2008)
10. V. Scherbahn, S. Nizamov, V.M. Mirsky, Toward ultrasensitive surface plasmon resonance sensors (2018)
11. P. Hausler, C. Genslein, C. Roth, T. Vitzthumecker, T. Hirsch, R. Bierl, Miniaturized surface plasmon resonance based sensor system, in *Proceedings of the 6th International Conference on Photonics, Optics and Laser Technology - Volume 1*, Photoptics (2018)
12. C. Rodriguez-Emmenegger, E. Brynda, T. Riedel, M. Houska, V. Šubr, A.B. Alles, E. Hasan, J.E. Gautrot, W.T. Huck, Polymer Brushes Showing Non-Fouling in Blood Plasma Challenge

- the Currently Accepted Design of Protein Resistant Surfaces. *Macromol. Rapid Commun.* **32**(13), 952–957 (2011)
13. H. Lísalová, E. Brynda, M. Houska, I. Visova, K. Mrkvova, X.C. Song, E. Gedeonova, F. Surman, T. Riedel, O. Pop-Georgievski, J. Homola, Ultralow-fouling behavior of biorecognition coatings based on carboxy-functional brushes of zwitterionic homo-and copolymers in blood plasma: functionalization matters. *Anal. Chem.* **89**(6), 3524–3531 (2017)
 14. J.W. Tomm, A. Jaeger, A. Bärwolff, T. Elsaesser, A. Gerhardt, J. Donecker, Aging properties of high power laser diode arrays analyzed by Fourier-transform photocurrent measurements. *Appl. Phys. Lett.* **71**(16), 2233–2235 (1997)
 15. J.H. Grassi, R.M. Georgiadis, Temperature-dependent refractive index determination from critical angle measurements: Implications for quantitative SPR sensing. *Anal. Chem.* **71**(19), 4392–4396 (1999)
 16. A.N. Naimushin, S.D. Soelberg, D.U. Bartholomew, J.L. Elkind, C.E. Furlong, A portable surface plasmon resonance (SPR) sensor system with temperature regulation. *Sens. Actuators B: Chem.* **96**(1–2), 253–260 (2003)
 17. O. Telezhnikova, J. Homola, New approach to spectroscopy of surface plasmons. *Opt. Lett.* **31**(22), 3339–3341 (2006)
 18. H. Šípová, M. Piliarik, M. Vala, K. Chadt, P. Adam, M. Bocková, K. Hegnerová, J. Homola, Portable surface plasmon resonance biosensor for detection of nucleic acids. *Procedia Eng.* **25**, 148–151 (2011)
 19. T. Brulé, G. Granger, N. Bukar, C. Deschênes-Rancourt, T. Havar, A.R. Schmitzer, R. Martel, J.F. Masson, A field-deployed surface plasmon resonance (SPR) sensor for RDX quantification in environmental waters. *Analyst* **142**(12), 2161–2168 (2017)
 20. B.N. Feltis, B.A. Sexton, F.L. Glenn, M.J. Best, M. Wilkins, T.J. Davis, A hand-held surface plasmon resonance biosensor for the detection of ricin and other biological agents. *Biosens. Bioelectron.* **23**(7), 1131–1136 (2008)
 21. K. Bremer, B. Roth, Fibre optic surface plasmon resonance sensor system designed for smart-phones. *Opt. Express* **23**(13), 17179–17184 (2015)
 22. Y. Liu, Q. Liu, S. Chen, F. Cheng, H. Wang, W. Peng, Surface plasmon resonance biosensor based on smart phone platforms. *Sci. Rep.* **5**, 12864 (2015)
 23. H. Guner, E. Ozgur, G. Kokturk, M. Celik, E. Esen, A.E. Topal, S. Ayas, Y. Uludag, C. Elbuken, A. Dana, A smartphone based surface plasmon resonance imaging (SPRi) platform for on-site biodetection. *Sens. Actuators B: Chem.* **239**, 571–577 (2017)
 24. J. Homola, I. Koudela, S.S. Yee, Surface plasmon resonance sensors based on diffraction gratings and prism couplers: sensitivity comparison. *Sens. Actuators B: Chem.* **54**(1–2), 16–24 (1999)
 25. J. Montague, Seriously? No kidding. Raspberry Pi, Arduino and other computers on open-source silicon boards are on the way for do-it-yourself monitoring—and even control. *Control* **30**(9), 34–40 (2017)
 26. M. Li, *Ranking Popular Deep Learning Libraries for Data Science* (2017). <https://blog.thedataincubator.com/2017/10/ranking-popular-deep-learning-libraries-for-data-science>
 27. M. Piliarik, J. Homola, Surface plasmon resonance (SPR) sensors: approaching their limits? *Opt. Express* **17**(19), 16505–16517 (2009)
 28. European Machine Vision Association, EMVA standard 1288, standard for characterization of image sensors and cameras. Release **3**, 1 (2016)
 29. H. Naumann, G. Schröder, M. Löffler-Mang, *Handbuch Bauelemente der Optik: Grundlagen, Werkstoffe, Geräte* (Carl Hanser Verlag GmbH Co KG, Messtechnik, 2014)
 30. F. Pedrotti, L. Pedrotti, W. Bausch, H. Schmidt, *Optik für Ingenieure* (Springer, Berlin Heidelberg, 2002)
 31. C.M. Keck, R.H. Müller, Size analysis of submicron particles by laser diffractometry—90% of the published measurements are false. *Int. J. Pharm.* **355**(1–2), 150–163 (2008)
 32. M.J. Weber, *Handbook of Optical Materials*, vol. 19 (CRC press, 2002)
 33. C.J. Lasance, A. Poppe (ed.), *Thermal Management for LED Applications* (Springer, Berlin, 2016)

34. G. Abbate, U. Bernini, E. Ragozzino, F. Somma, The temperature dependence of the refractive index of water. *J. Phys. D Appl. Phys.* **11**(8), 1167 (1978)
35. J.R. Janesick, *Scientific Charge-Coupled Devices*, vol. 83 (SPIE press, 2001)
36. L. Niu, N. Zhang, H. Liu, X. Zhou, W. Knoll, Integrating plasmonic diagnostics and microfluidics. *Biomicrofluidics* **9**(5), 052611 (2015)
37. Y. Song, D. Cheng, L. Zhao (eds.) *Microfluidics: Fundamentals, Devices, and Applications* (Wiley, 2018)
38. L. da Fontoura Costa, R.M. Cesar, *Shape Classification and Analysis: Theory and Practice* (CRC Press, Inc, 2009)
39. T. Klinger, *Image Processing with LabVIEW and IMAQ Vision* (Prentice Hall Professional, 2003)
40. A. Zybin, D. Boecker, V.M. Mirsky, K. Niemax, Enhancement of the detection power of surface plasmon resonance measurements by optimization of the reflection angle. *Anal. Chem.* **79**(11), 4233–4236 (2007)
41. S. Nizamov, V. Scherbahn, V.M. Mirsky, Self-referencing SPR-sensor based on integral measurements of light intensity reflected by arbitrarily distributed sensing and referencing spots. *Sens. Actuators B: Chem.* **207**, 740–747 (2015)
42. S. Nizamov, V.M. Mirsky, Self-referencing SPR-biosensors based on penetration difference of evanescent waves. *Biosens. Bioelectron.* **28**(1), 263–269 (2011)
43. A.K. Sharma, B.D. Gupta, On the performance of different bimetallic combinations in surface plasmon resonance based fiber optic sensors. *J. Appl. Phys.* **101**(9), 093111 (2007)
44. B.H. Ong, X. Yuan, S.C. Tjin, J. Zhang, H.M. Ng, Optimised film thickness for maximum evanescent field enhancement of a bimetallic film surface plasmon resonance biosensor. *Sens. Actuators B: Chem.* **114**(2), 1028–1034 (2006)
45. V. Švorčík, P. Slepíčka, J. Švorčíková, M. Špírková, J. Zehentner, V. Hnatowicz, Characterization of evaporated and sputtered thin Au layers on poly (ethylene terephthalate). *J. Appl. Polym. Sci.* **99**(4), 1698–1704 (2006)
46. B.A. Sexton, B.N. Feltis, T.J. Davis, Characterisation of gold surface plasmon resonance sensor substrates. *Sens. Actuators A* **141**(2), 471–475 (2008)
47. S.A. Zynio, A.V. Samoylov, E.R. Surovtseva, V.M. Mirsky, Y.M. Shirshov, Bimetallic layers increase sensitivity of affinity sensors based on surface plasmon resonance. *Sensors* **2**(2), 62–70 (2002)
48. D.V. Nesterenko, Z. Sekkat, Surface plasmon sensing with different metals in single and double layer configurations. *Appl. Opt.* **51**(27), 6673–6682 (2012)
49. L. Pang, G.M. Hwang, B. Slutsky, Y. Fainman, Spectral sensitivity of two-dimensional nanohole array surface plasmon polariton resonance sensor. *Appl. Phys. Lett.* **91**(12), 123112 (2007)
50. J.F. Masson, M.P. Murray-Méthot, L.S. Live, Nanohole arrays in chemical analysis: manufacturing methods and applications. *Analyst* **135**(7), 1483–1489 (2010)
51. C. Genslein, P. Hausler, E.M. Kirchner, R. Bierl, A.J. Baeumner, T. Hirsch, Detection of small molecules with surface plasmon resonance by synergistic plasmonic effects of nanostructured surfaces and graphene, in *Plasmonics in Biology and Medicine XIV*, vol. 10080, p. 100800F. International Society for Optics and Photonics (2017)
52. C. Genslein, P. Hausler, E.M. Kirchner, R. Bierl, A.J. Baeumner, T. Hirsch, Graphene-enhanced plasmonic nanohole arrays for environmental sensing in aqueous samples. *Beilstein J. Nanotechnol.* **7**, 1564 (2016)
53. M.C. Estevez, M.A. Otte, B. Sepulveda, L.M. Lechuga, Trends and challenges of refractometric nanoplasmonic biosensors: A review. *Anal. Chim. Acta* **806**, 55–73 (2014)
54. D. Herbert, Batch to Continuous. *Control* **22**(9), 48–55 (2009)
55. M. Heim, Chemie im Blut. *Zeit Online* (2018). <https://www.zeit.de/gesellschaft/zeitgeschehen/2018-01/trinkwasser-chemikalien-bayern-altoetting-ueberland/komplettansicht>

Chapter 9

Photonics-Enhanced Image-Detection Sensing of Multiphase Flows



Sergio L. Carrasco-Ortiz, Eduardo Valero, Maria Morant and Roberto Llorente

Abstract This chapter describes a photonic sensor system based on laser excitation and CMOS array image capture altogether advanced digital signal processing algorithms. The photonic sensor targets the detection and characterization of cavitation bubbles in multiphase water flows. This sensor finds application in areas where a multiphase water flow is produced, by example bubbling water column reactors, turbine impellers, marine screw and pump-jet propellers where cavitation can be produced, and water-air mixing volumes in dam intakes and spillways in hydroelectric energy generation plants. The photonic sensor comprises an image capture CMOS array with a polymeric tunable optical lens which digitises an area illuminated by a laser diode operating at wavelength 532 nm. This approach permits high-contrast acquisition independent of external lighting conditions. Ad hoc signal processing algorithms are applied on the digitised image in order to evaluate the statistical distribution of bubble size, shape, speed and concentration inside the multiphase flow. Experimental demonstration of the developed sensor indicates its proper operation, being capable of a complete statistical bubble characterization in a water column at 0.01 and 0.05 MPa pressure levels. The performance of different computational methods, including Optical Flow, SIFT and SURF, has been also evaluated in the experimental work for comparison of the underlying image processing algorithms.

9.1 Introduction

A multiphase flow is defined as a simultaneous flux comprising two or more phases or components with an explicit interface between them [1]. A basic example of a two-phase flow is the one formed by gas and a liquid, being the most common example a bubble flow [2]. The characterization of multiphase flows requires the detection of bubbles and the statistical characterization of the bubble key parameters: size, shape, speed and fractional proportion [3].

S. L. Carrasco-Ortiz (✉) · E. Valero · M. Morant · R. Llorente
Nanophotonics Technology Center, Universitat Politècnica de València, Valencia, Spain
e-mail: sercaror@ntc.upv.es

© Springer Nature Switzerland AG 2019
P. Ribeiro and M. Raposo (eds.), *Optics, Photonics and Laser Technology 2018*, Springer Series in Optical Sciences 223,
https://doi.org/10.1007/978-3-030-30113-2_9

Adequate characterization of multiphase flows plays a crucial role in many industrial and medical fields. Characterization of bubble flows is a preliminary step for monitoring the behaviour of gas/liquid two-phase flows, which is of utmost importance [2]. Two examples of monitoring applications of two-phase flows can be found in cavitation detection in laser atherectomy [4], in air bubble detection in extracorporeal blood circuits of haemodialysis [5]. In this latter application, if a certain amount of air is pumped into the patient, it may cause life-threatening coagulation and haematomas [5]. Other applications include industrial bubble column reactor monitoring [1]. In this case, it is extremely important to know the properties of the bubbles including its size, shape and velocity for the design and functional monitoring of the reactor [1]. In both applications, image processing of bubbles has been proposed in the state-of-the-art to calculate the number of bubbles and cumulative volume in a given flow [6, 7].

Considering hydraulic applications, which are the main scope of this chapter, a large number of bubbles in a flow can indicate the presence of cavitation. Cavitation is a phenomenon that generates bubbles, which explode in the blade surface due to the pressure decrease generated by an excessive blade angular velocity [8]. Areas with cavitation phenomenon are usually involved in highly turbulent flows as it occurs in vortex emissions produced by the torch effect represented in Fig. 9.1a [9]. In particular, in hydraulic structures, the explosion of bubbles causes erosion on the surface of dam spillways. This erosion reduces the useful life of the structure [10] as it can produce severe damages in the structure itself as it is depicted in Fig. 9.1b in an example of the Nagarjunasagar Dam located in India [11]. Moreover, in devices such as pumps, hydraulic turbines, pump-jets and screws in the naval field (installed on ships and submarines), cavitation could cause noise, vibrations, permanent screw damage and propulsion efficiency losses [12, 13].

Due to the compressibility of bubble flows, air is capable of absorbing (until a certain point) the impact of bubble shocks in cavitation processes [14]. However, if

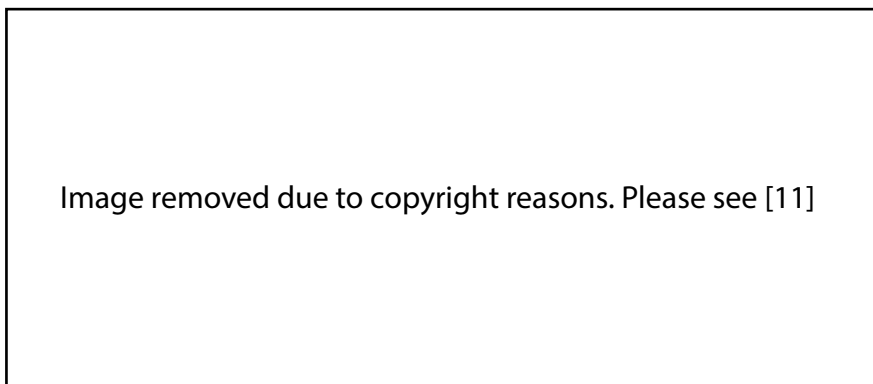


Fig. 9.1 **a** Cavitation in a turbine impeller, **b** cavitation damage (after several floods in 2009) on spillway profile of Nagarjunasagar Dam [11]

the bubble flow is not properly monitored the cavitation phenomenon can lead to the above-mentioned damages in the structure. The aeration of the flow changes its dissipative behaviour—as provides less friction with the contours and increased vorticity—[15], which converts the concentration of air in a variable of crucial interest. The level of bubble concentration modifies the macroscopic properties of the flow, increasing its draft and decreasing its velocity. In addition, significant differences have been found [16] in the behaviour of the turbulence due to the effect of the dispersed phase (in this case, air) on the carrier phase (in our case, water).

The most common solution to cavitation in hydraulic structures is the design of hydro-dynamic profiles that avoid big pressure gradients [17, 18]. However, this accurate design is not always possible in practice and cavitation may occur at relatively slow speeds. For these reasons, other sort of solutions are necessary to detect and avoid the cavitation phenomenon. The employment of specific coating materials has been also proposed in the past to improve the resistant properties [19]. Acoustic detection of the cavitation bubbles implosion has been also evaluated in the literature [20], where incipient cavitation is heard as intermittent crackling while severe cavitation becomes a steady hiss increasing in volume [21, 22]. However, due to the chaotic rise of the bubbles along the flow in both number and size, it is not possible to predict the emitted noise to get a full characterization [21]. Other techniques for characterizing bubbles are based on energy measurement—employing charge-coupled devices (CCD) or hydrophones [23]—, electric impedance measurement [24], or based on contact-probes that measure if there is air or water in contact with the probe [25]. All these methods are invasive, and they will modify the properties of the hydrodynamic multiphase flow itself, such as bubble shape and velocity. To avoid this, we propose a non-invasive photonic sensor based on an image capture sensor and a diode laser that does not interfere with the multiphase flow. The diode laser is employed to excite the bubbling water column in order to increase the contrast of the high-density number of bubbles in the flow. This non-invasive method enables monitoring the flow behaviour taking high-speed images of the flow and post-processing the captured images.

In this chapter, we describe the optical sensor and the developed associated digital processing algorithms for the characterization of the multiphase flows. The work is focused on the detection and characterization of bubbles in a multiphase flow water column to obtain representative information about the air concentration and number of bubbles per sensed area, including the dimension and velocity of the detected bubbles. A potentially high bubble-density is considered in the experimental evaluation in order to address the proposed scenarios.

The developed image processing algorithms permit the estimation of bubble speed, bubble size and also shape characterization in terms of equivalent area, diameter, major half-axis, minor half-axis and perimeter from the captured data.

This chapter is structured as follows: in Sect. 9.2, the architecture of the proposed photonic sensor is described including the main photonic subsystems and the associated processing algorithms. In Sect. 9.3, the experimental bubble characterization results for both experiments, are shown and described in terms of bubbles' concentration, bubbles' shape and bubbles' speed. Finally, in Sect. 9.4, the main conclusions of this work are summarized.

9.2 Laser-CMOS Sensor Architecture

9.2.1 Photonic Subsystems

Figure 9.2 shows the conceptual scheme of the laboratory setup employed in this experimental evaluation. We can observe a block diagram of the main components that comprise the proposed photonic sensor for the detection and characterization of bubbles in a multiphase flow. In this evaluation, the sensor is applied to a bubbly water column generated at the laboratory inside a transparent tank with two different cylindrical diffusers with length of $L_{\text{diff}} = 26.50$ mm (one of them with an unknown porous diameter and another one with a diameter $\varnothing_{\text{diff}} = 12.10$ mm). The cylindrical diffusor is connected with a flexible plastic tube ($\varnothing_{\text{in}} = 5.10$ mm, $\varnothing_{\text{out}} = 8.10$ mm) to a void pump configurable from 0 to 1 MPa.

The laser-assisted sensor is based on a camera that comprises a complementary metal-oxide-semiconductor (CMOS) sensor, a fixed glass lens and a tunable polymeric lens controlled remotely with a driver. The illumination of the bubbles is achieved with a laser diode beam dispersed by a negative lens. The computing is performed by a computer with an 8-core CPU and with a 8 Gb RAM memory.

The primary functions of each opto-electronic component depicted in Fig. 9.2 can be summarized in:

- Monochrome CMOS sensor**, which is responsible device of data acquisition: Each one of the pixels of the CMOS incorporates the necessary electronics to convert its own electric charge into potential difference independently, so that it is possible to access the information captured in a certain area. Thanks to that, CMOS sensors have smaller dimensions and lower energy consumption than others as Coupled-Charged Device (CCD) sensors. The sensor is configured remotely in terms of gain, resolution, frames per second (fps), recording time and exposure time (μs). In these experiments, the sensor is configured according to the parameters included in Table 9.1.

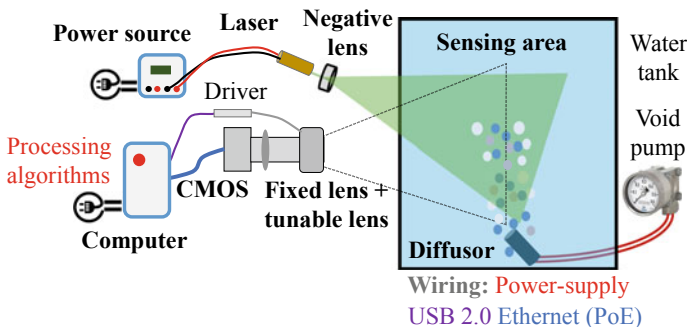
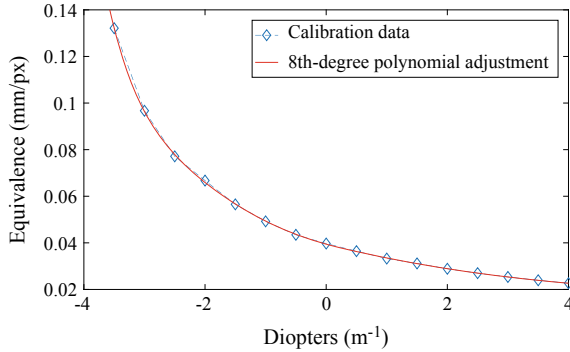


Fig. 9.2 Conceptual scheme of the laser-assisted CMOS sensor architecture developed at the laboratory [3]

Table 9.1 CMOS sensor configuration

CMOS parameters	Value
Gain	5
Resolution	1280 × 1024
Frames per second (fps)	200
Recording time (s)	5
Exposure time (μs)	200

Fig. 9.3 Calibration equivalence of mm/px versus dioptres of the tunable lens system



- **Fixed lens**, with a focal length of $f = 25$ mm is the interface between the tunable lens and the sensor focusing the image on the camera sensor.
- **Tunable lens**, which is employed for lens system adjustment: This lens, made out of polymeric materials, is capable of varying its dioptre number between -4 and $+4$ dpt to focus in different depth planes depending on how positive or how negative is the electric current supplied. By means of algorithms that set the tunable lens current, we are able to control the lens system. Those algorithms are necessary for the calculation of the real size. These algorithms calculate the equivalence between the image size in terms of pixels (px) and the real length in mm. Figure 9.3 shows the calibration curve of the tunable lens system in terms of mm/px versus dioptres.

In this work, the refractive surface between the two media (air and water) is plane, which means that the refraction does not imply any increase in size. Applying a polynomial adjustment of 8th degree to the curve represented in the graph of in Fig. 9.3, (9.1) is obtained with the coefficients included in Table 9.2.

$$Eq = \sum_{k=0}^8 c_k \cdot dpt^k \tag{9.1}$$

- **Laser diode**, which illuminates the bubbles volume under analysis increasing the contrast and shadowing of the borders. A tunable output power laser with up to 80 mW operating at a wavelength of $\lambda = 532$ nm is employed. The concrete wavelength was selected to match the quantum efficiency of the CMOS sensor.

Table 9.2 8th degree polynomial adjustment coefficients show in Fig. 9.3 [3]

Coefficients	Value
c_0	0.0405
c_1	-8.01×10^{-3}
c_2	2.62×10^{-3}
c_3	-1.10×10^{-3}
c_4	-7.33×10^{-5}
c_5	1.22×10^{-4}
c_6	2.54×10^{-6}
c_7	-8.16×10^{-6}
c_8	9.07×10^{-7}

Figure 9.4 shows the efficiency of the CMOS sensor for different wavelengths ranging from 400 to 1000 nm. The optimal quantic efficiency (over 68%) of the CMOS sensor is obtained at wavelength $\lambda = 532$ nm.

The geometrical laser ray has a roundness $O > 90\%$, a divergence $\theta > 1.2$ mrad and an initial diameter $\varnothing_0 = 1.5$ mm. The output power of the laser can be configured depending on the voltage/intensity provided ranging from 2.3 to 2.9 V in voltage and up to 350 mA in current. Following the manufacturer’s recommendations, in order to ensure the output power, a 10-s guard period is set before capturing any information after switching the laser.

- **Negative lens**, which is included after the laser to diverge the optical beam. If the laser is too directive it could produce reflection or sparkles in the illuminated bubbles, which would cause saturation in the data captured with the CMOS sensor. To avoid this effect, the laser beam is diverged with a negative lens with the physical specifications depicted in Table 9.3.

Fig. 9.4 CMOS quantum efficiency

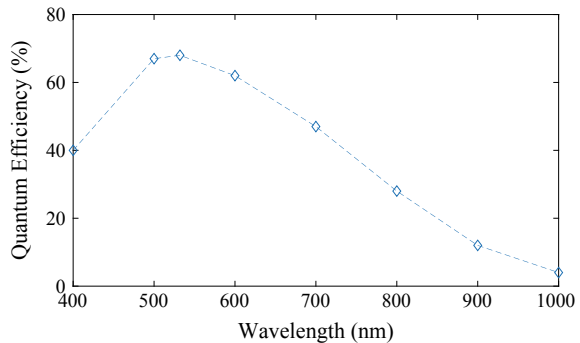


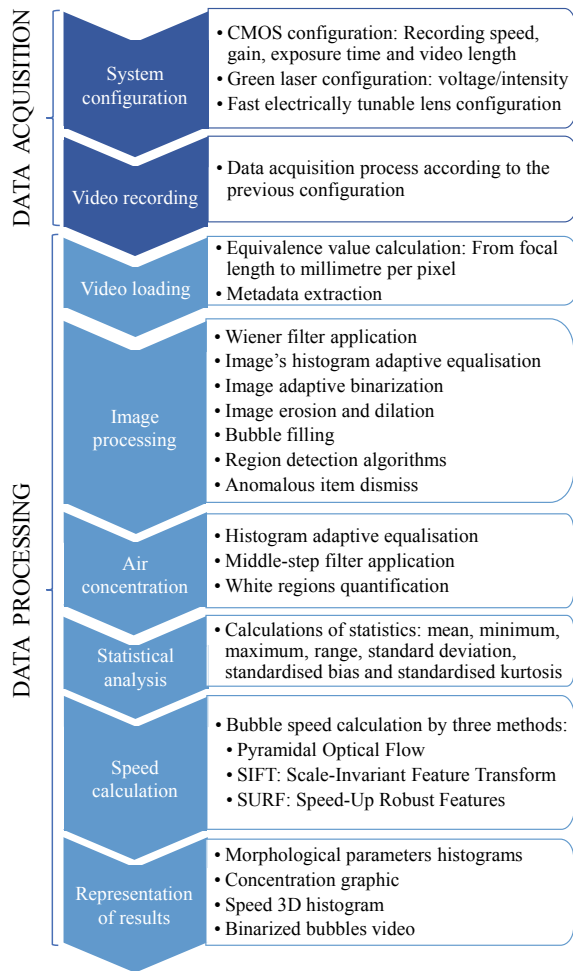
Table 9.3 Lens description

Physical parameters	Description
Glass type	NBK-7
Shape	Plane-Concave
Diameter (mm)	50.8
Focal distance (mm)	-75

9.2.2 Data Acquisition and Digital Signal Processing

Figure 9.5 summarizes the main steps defined in the developed digital signal processing (DSP) algorithms to fully characterize the bubble flow. The DSP algorithms

Fig. 9.5 Data acquisition and processing diagram [3]



are programmed in MatLab to configure the data acquisition and post-processing the captured images. As a first step, the sensor is configured remotely via C commands including the CMOS configuration, the tunable lens focus and the output power of the laser diode.

In the first step of System Configuration depicted in Fig. 9.5, the CMOS sensor is configured in terms of:

- **Gain correction** employed to homogenise the CMOS sensibility. Implementing dead pixel correction removes empty pixels. This correction consists in storing the average value of the adjacent pixels in the black pixel. This process improves the detection area of the images in post-processing, however it may induce errors in the image's adaptive histogram equalisation that should be taken into account.
- **Exposure time** used to configure the shutter speed. The exposure time adjusts the amount of incident light entering to the sensor. The value of the exposure time should be adequate to avoid blurred images.
- **Recording speed:** The readout speed indicates how many frames-per-second are captured by the camera. In these experiments, two recordings are configured with different recording speeds. In first place, we capture images at a speed of 100 fps (i.e. each frame is captured at 1/100 s or each 10 ms) to evaluate the air concentration and morphology of the detected bubbles. In second place (and immediately after the previous capture), the recording speed is set to 1110 fps (i.e. 1/1110 or 0.9 ms) for the estimation of the velocity of the bubbles.

Once the sequence of images is recorded, according to the focus configuration of the lens, the equivalence of mm/px is obtain using the calibration curve depicted previously in Fig. 9.2. The captured images are loaded in MatLab and metadata information is extracted.

The first step in the image post-processing is the recognition of any object carried out by the discrimination of what is seen regarding the ideal concept of the object to be identified. To achieve this, it is necessary to transform the images acquired in greyscale—pixels whose values go from 0 to 1, being 0 a black pixel and being 1 a white pixel—into images only in black and white. This process is known as binarization as only 0 or 1 values remain in the image—0 for black pixels and 1 for white pixels.

To minimize the interclass variance between white and black pixels, it is necessary to have some ambient lighting during the capture of the information. An accurate conversion from a grey image to a binarized image is essential for the application of region detection algorithms. The laser diode illumination provided by the proposed optical sensor increases the contrast of the bubbles in the flow and enhances the binarization process.

To carry out the identification of circular/elliptical forms, some extra filtering and processing should be implemented in the post-processing of the captured images. The filters and processes applied to the images are integrated in C/Fortran language. The different image processing steps applied to the images captured from the bubbly water column can be summarized as:

- **Wiener Filter [26]:** One of the main problems generated by the application of the histogram equalization algorithms is the appearance of extra noise when turbidity. This is a random and non-controllable variable. Due to the effect of turbidity in the image, it will be essential to eliminate as much noise as possible so that, once the image is binarized, the minimum number of “corpuscles” may appear. This step is very important in the post-processing of the images because these globules produced by noise could be confused with bubbles. To avoid this, a mid-pass filter is implemented as it can be observed in Fig. 9.6. This non-linear operation is used to eliminate noise in an image while maintaining the edges of the items present in the image, as it can be confirmed in the output image shown in Fig. 9.6b.
- **Histogram Adaptive Equalization:** The *contrast limited adaptive histogram* [27] is a generalization of the histogram equalization and the equalization of the adaptive histogram. This technique does not work in the whole image (as in the case of the equalization of the histogram), but it works in small areas of the image called boxes. In this way, the contrast of each box is improved (see Fig. 9.7). The adjacent boxes are combined by the use of bilinear interpolation to eliminate artificial boundaries between induced boxes. The contrast (particularly in homo-

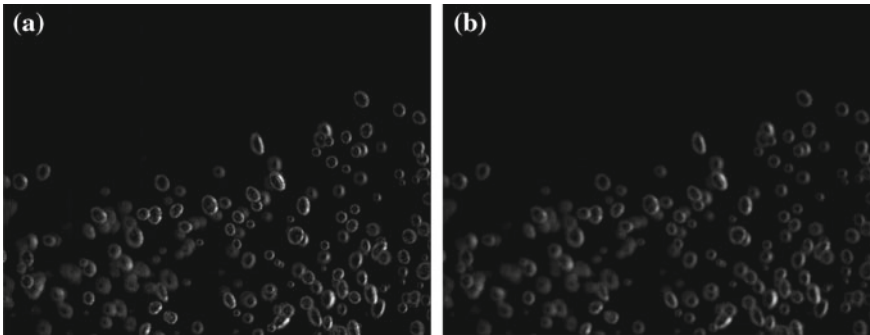


Fig. 9.6 a Raw input image, b Wiener filter output image

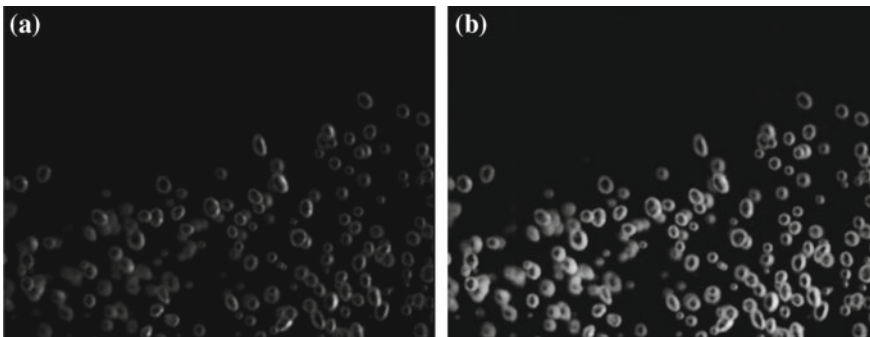


Fig. 9.7 a Wiener filter image, b output image from histogram adaptive equalisation

geneous regions) is limited to avoid the amplification of unwanted information such as existing noise. To prevent saturation, a contrast factor β is used, which prevents oversaturation of the image specifically in homogeneous areas. These areas are characterized by high peaks in the histogram of the box of a particular image where several pixels correspond to the same grey range.

- **Adaptive Binarization:** The binarization process [28] of an image is based on the comparison of grey levels present in the image with a given threshold. If the grey level of the pixel under evaluation (originally in greyscale) is smaller than the given threshold, a value of 0 (corresponding to black) is assigned to that pixel, whereas, if the value is higher, a value of 1 (white) is assigned. The outcome of this process is a black and white image—hence the name of binarization—as shown in Fig. 9.8.

As it was discussed before, the binarization process is a critical step in the image post-processing, since, if it is not done correctly, the bubbles will not be displayed correctly (with pixels values set to 1) and, later, when applying the circular/elliptical search process, we won't be able to identify the bubbles in the image. For this reason, it is very important to use the optimal threshold value for the binarization of the captured images.

- **Image Erosion:** The goal of the erosion filter is to eliminate the noise that appears in a binarized image such as random points or small blocks of white pixels. The input signals for the erosion process are the images already binarized and a structural object called *strel*. The *strel* object [29] represents a morphologically flat structural element (disk, line, square, rectangle, etc.) of size R that is used to both the erosion and dilatation operations. An example of the output from the erosion filter is shown in Fig. 9.9b.
- **Image Dilation:** When applying the erosion filter, there will also be a loss of white pixels that could be essential for the detection of bubbles. Therefore, dilation filter is implemented to recover the white pixels lost in the bubbles using the same *strel* object as defined previously. An example of dilation filter recovery is shown in Fig. 9.10.

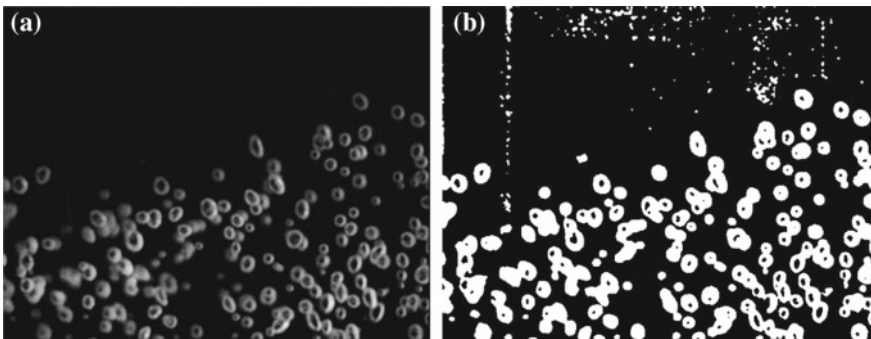


Fig. 9.8 a Image from histogram adaptive equalisation, b adaptive binarization output image

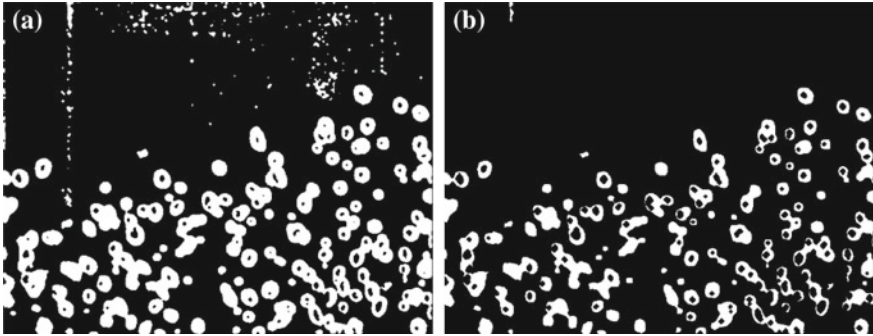


Fig. 9.9 a Image adaptive binarization image, b image erosion

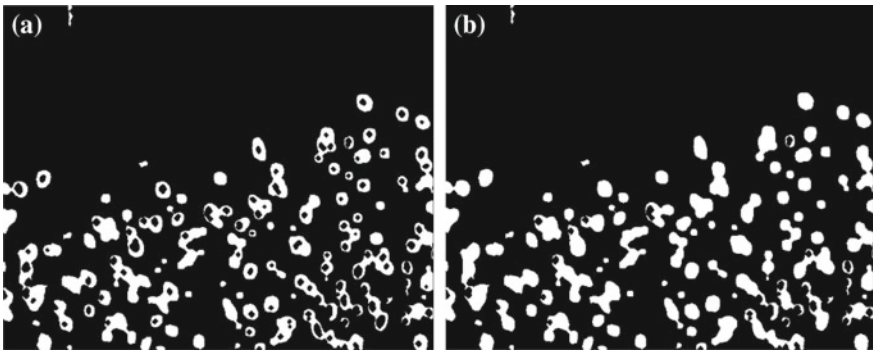


Fig. 9.10 a Input image coming from the erosion filter, b image dilation

The procedure of eroding first and then dilating is known as “opening” [30]. With this method, the previously removed noise will not appear again in the image if the *strel* object is defined properly to completely remove the noise.

- **Bubble Filling:** The main objective of filling the bubbles is to reduce the algorithms’ compilation time. The holes (black pixels totally surrounded by white pixels) found in the image are filled with white pixels. Implementing this filling, we avoid wasting time looking for circular or elliptical shapes inside larger ones. An example of this output image is depicted in Fig. 9.11b.
- **Anomalous Item Dismisses:** After the processing of images with bubbles, it is possible that abnormal regions appear during binarization. An example of this anomalous item is included in Fig. 9.12a.

To avoid these anomalous items, not all the population of the binarized regions in a frame will be considered as bubbles for the statistical treatment of morphological parameters of the bubbles. The anomalous items comprise strange shape regions obtained after binarized or product of binarization of two or more bubbles that could appear together as a white region. In this case, discard algorithms are implemented

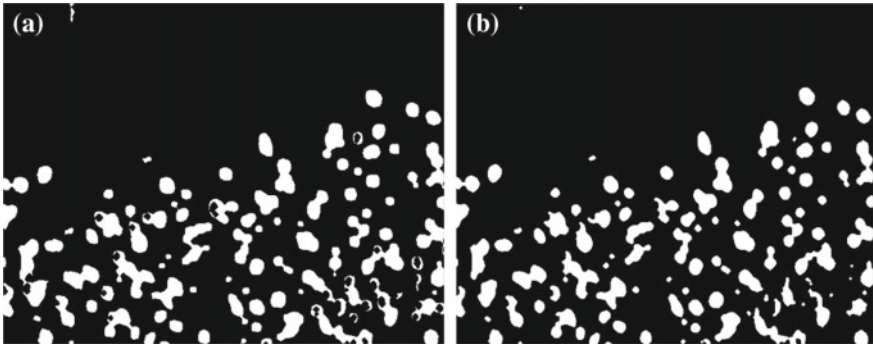


Fig. 9.11 a Dilated image, b bubble filling image

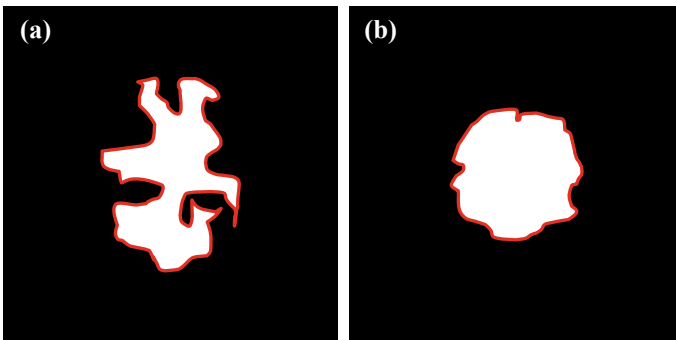


Fig. 9.12 a Anomalous region, b bubble region

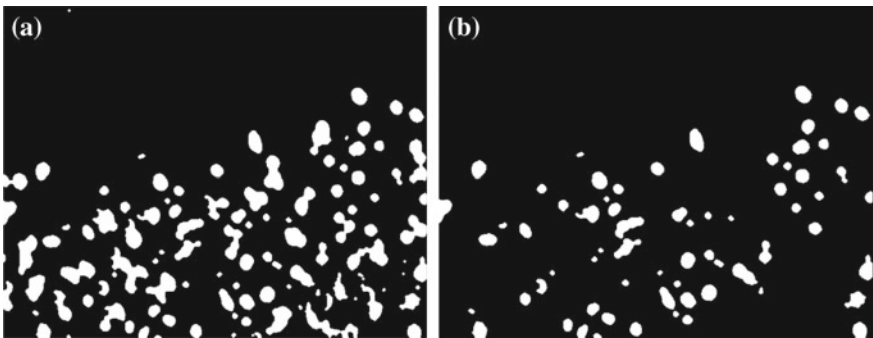


Fig. 9.13 a Dilated image, b clean image after anomalous items dismiss

in our post-processing software based on eccentricity conditions and solidity of the bubbles. Figure 9.13 shows the final result of the post-processed bubbles to

be analysed morphologically after dismissing the anomalous regions of the dilated image.

In this step, it is essential that there is no blurring in the image. In addition, the lighting conditions should be optimal to capture the bubbles correctly and recognize the shapes in their entirety without becoming confused with each other.

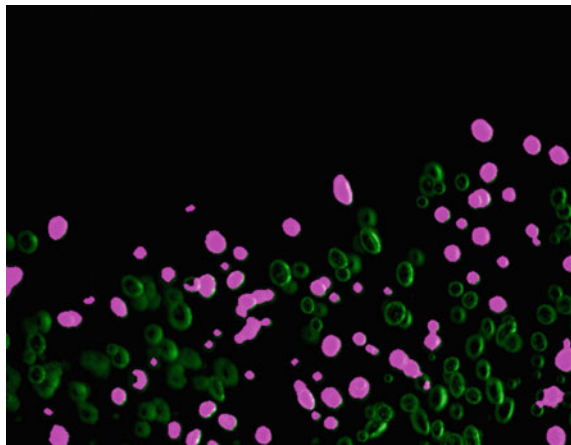
Moreover, this process is not done only to omit abnormal regions; it is also used to debug bubbles that being too close each other, after processing, compose a false new region bigger than the original bubble. During the validation experiments, bubbles greater than 1.5 times its statistical deviation were debugged.

Figure 9.14 shows a frame example where the bubbles highlighted are selected for the statistical and morphological analysis.

Next, after the statistical analysis of the detected bubbles, the bubbles' speed is calculated. In this work, we performed an exhaustive state-of-the-art study of the methods available for the calculation of the velocity profile of moving objects. In particular, in these experiments we compare three evaluation methods based on algorithms available in the MatLab *Image Processing toolbox*:

- **Optical Flow:** This method is based on the distribution of apparent velocities of brightness patterns. The Optical flow algorithm arises from the relative movement between the objects and the observer, thus, the movement of the observer is not taken into account. The Horn-Schunck estimation method [31] is one of the classical optical flow algorithms due to its simplicity and good results. This algorithm is based on a differential computation technique that requires a conservation of brightness with a global smoothing to obtain an estimation of the velocity fields. The velocity estimation is composed of a two-dimensional vector. The optical flow is computed at a given point taking into account the adjacent points considering that the brightness has only one component and the speed has two components.
- **Scale-Invariant Feature Transform (SIFT):** This method consists in the extraction of relevant characteristics that can be used for object recognition or motion

Fig. 9.14 Example of processed frame with highlighted bubbles for the statistical analysis



detection [32]. The most important aspect of the SIFT method is the high number of milestones that are calculated on the image for the recognition of moving objects. For example, an image with a size of 500×500 pixels can have around 2000 milestones for the recognition of objects, although this number can vary depending on the content of the images themselves. The computation steps carried out by this method for the identification of objects and the calculation of the speed are:

- *Scale-space Extrema Detection*: In this first step the algorithm searches over all image areas, extremes, scales and edges. It is implemented efficiently by using the Gauss increment function to identify potential points of interest that remain invariant in the images.
 - *Keypoint Localization*: For each of the candidate areas, invariant points are selected based on stability measures.
 - *Orientation Assignment*: One or more-pixel orientations are assigned to each of the key points designated in the candidate area of the image. All subsequent operations performed on the image are based on the assigned orientation, scale and location for each of the characteristic points that remain invariant.
 - *Keypoint Descriptor*: The gradients of the local image are measured on the scale selected in the region around each of the key points.
- **Speeded-Up Robust Features (SURF)**: This method extracts information from the image to detect items, realize 3D scenes and track objects inside a video in movement [33]. The SURF algorithm searches for this information in discrete images by implementing these three steps:
 1. *Localization of Points of Interest*: These points (also known as optimal coordinates for the calculation of speeds) are selected in different areas of the image, such as corners, T-junctions and small circles. The most important property of the points of interest is its repetitiveness. Thus, the detector relies on finding the same physical points of interest in different images. To obtain these points of interest four steps are followed:
 - 1.1 Integral of the images.
 - 1.2 Identification of points of interest based on the Hessian matrix.
 - 1.3 Space-scale representation.
 - 1.4 Location of the point of interest.
 2. *Descriptors definition*: The points of interest described in the previous step have a characteristic vector called descriptor. This descriptor must be distinctive and at the same time robust with regard to noise, geometric deformations and light.
 3. *Marking*: In this step, the descriptors are marked between the different images. This marking is based on the difference of distances between vectors in more than two images. The calculation of the marking has a direct cost in computing time. The determination of the descriptors and their marking is carried out in the following steps:

- 3.1 Assignment of the orientation.
- 3.2 Descriptor based on the sum of Haar wavelet responses.
- 3.3 Rapid indexing for marking descriptors.

9.3 Experimental Bubble Characterization Results

The validation of the developed processing algorithms using the proposed sensor architecture is performed at the laboratory with bubbly water columns within a transparent tank. The experimental setup was depicted previously in Fig. 9.2. In particular, two experiments were carried out: the first one with the void pump set to 0.01 MPa (experiment #1) and a second one with 0.05 MPa (experiment #2) in order to generate a turbulent bubbly flow in a water column.

9.3.1 Air Bubble Concentration Calculation

The concentration of bubbles, understood as the amount of undissolved air (bubbles) detected by the sensor, is carried out by the application of for-type loops. The data obtained from all the white regions generated after the binarization of the image is used for the calculation of the concentration of air present in the image. Figure 9.15 depicts an example of a processed binarized image for the calculation of the air bubble concentration.

The sensor has been programmed to quantify the percentage of undissolved air present in each of the images captured by the CMOS sensor. In this case, all the bubbles are considered in the analysis, including also the bubbles dismissed from the statistical analysis for reasons of grouping, solidity and/or eccentricity. The devel-

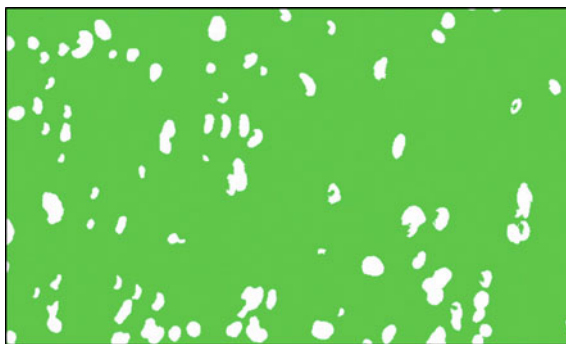
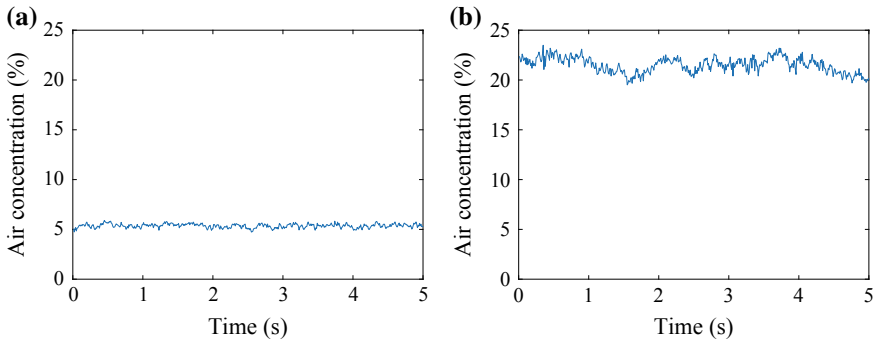


Fig. 9.15 Example of a binary image of multiphase flow used for the calculation of air bubble concentration in a given area

Table 9.4 Experimental quantification of bubbles detected in the water column

Experiment	Number of detected bubbles	Number of analysed bubbles for morphology statistical study
#1	737	571
#2	3547	1583

**Fig. 9.16** Non-diluted air concentration detected in: **a** experiment #1, **b** experiment #2

oped algorithms are able to report the total number of bubbles detected and the number of bubbles that can be analysed by the region detection algorithms. A summary of the processed data for both experiments is included in Table 9.4.

In the graphs shown in Fig. 9.16, 500 data points of concentration are represented in a capturing time of 5 s. Considering both the minimum and maximum peaks, in Fig. 9.16a the air concentration varies from 5.11 to 6.98%, while in Fig. 9.16b it fluctuates from 19.51 to 23.49%.

Comparing both graphs, Fig. 9.16a shows an average air concentration in the sensing zone of 6.03%, while Fig. 9.16b shows an average air concentration of 21.50% for the sensing area.

9.3.2 Analysis of Bubbles' Shape

In the case of ideally spherical bubbles, the detection and determination procedure would be as simple as the application of the Hough transform in the binarized image (after the post-processing with the filters described previously in Sect. 9.2.2). However, ideally spherical bubbles only occur under certain hydraulic conditions. For this reason, algorithms for the detection of regions are implemented to analyse different parameters, such as:

- *Surface*: Returns the area of the detected bubbles in number of pixels of that region. This value is then converted to a real length (mm^2) by means of the equivalence value obtained previously.
- *Equivalent Diameter*: Returns a scalar value that indicated the equivalent diameter of a circle with the same area as the blank detected region.
- *Major Half-axis*: Returns a scalar value of the length expressed in pixels of the major half-axis of the ellipse that has the same normalized central moment as the blank region.
- *Minor Half-axis*: Returns a scalar of the length in pixels of the minor half-axis of the ellipse that has the same normalized central moment as the blank region.
- *Perimeter*: Returns a scalar equivalent to the number of pixels that wrap around the white region obtained after processing the image.

These parameters are subject to the filtering of discarding anomalous objects as described previously in Sect. 9.2.2. In order to minimize the error, the anomalous regions and the aggrupation of bubbles are not considered for the morphological analysis.

Figures 9.17 and 9.18 depict the frequency histograms obtained with the post-processing algorithms; being the value of the sum of its bars equal to 1 represented in 25 columns. The type of histogram, as well as the number of columns in which the sensor data are represented can be modified in the developed algorithms. These histograms represent the distribution of the morphological parameters for experiment #1 and experiment #2.

By means of statistical analysis of the sensed data, including the mode, mean, minimum, maximum and typical deviation, the morphological shape of the bubbles present in the water column flow is evaluated. The deviations in shape parameters appear due to the non-uniform porous diameter of the diffusor that generates bubbles with different sizes as expected in a real-case scenario. Table 9.5 and Table 9.6 summarize the statistical results obtained in experiment #1 with raw data and debugged data, respectively. Whereas Table 9.7 and Table 9.8 summarize the statistical analysis, results obtained in experiment #2 with raw data and debugged data, respectively.

9.3.3 Bubbles' Speed Calculation

One of the most interesting calculations obtained with the proposed sensor is the estimated velocity at which the bubbles move in the flow. This estimation is calculated with different algorithms including SURF, SIFT and Optical Flow Pyramidal methods, described previously in Sect. 9.2.2. Figure 9.19 shows the result of experiments #1 and #2 for the velocity of the bubbles. The results are represented in three-dimensional frequency histograms to compare the results obtained with each method.

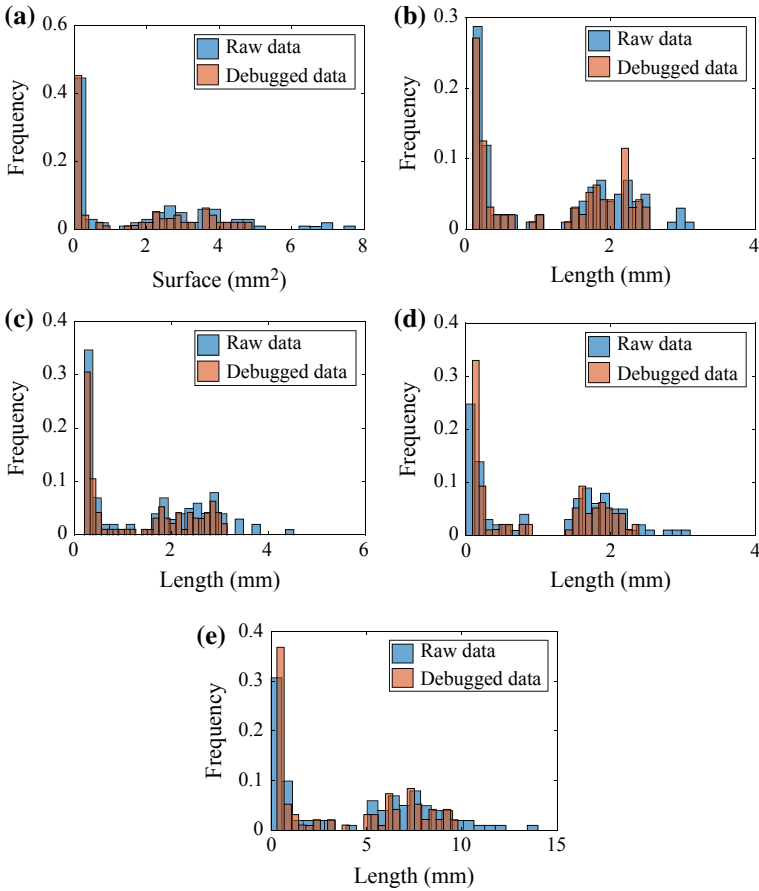


Fig. 9.17 Bubble shape histograms from experiment #1: **a** area, **b** equivalent diameter, **c** major half-axis, **d** minor half-axis, **e** perimeter. *Raw data* Data obtained without considering bubble superposition. *Debugged data* Data obtained after debugging 1.5σ times its shape parameter

The mean and the mode of the estimated bubbles’ speed obtained with experiments are summarised in Table 9.9.

As it can be observed in Table 9.8, all the speed estimation algorithms for experiment #1 have a mode of 0.3 m/s while in experiment #2 the mode is of 0.4 m/s. The mean of the velocity estimation differs between algorithms. The SIFT estimation has a mean slightly higher than Optical Flow and SURF methods but only in a 5% for experiment #2. However, for experiment #1, the mean obtained with SIFT estimation is higher in 70%. This indicates that SIFT method less suitable than SURF and Optical Flow methods for the calculation of low levels of bubbles’ speed in a water flow.

As the Optical Flow method performs the calculation of movement in each pixel based on the constancy of the light intensity, and such as is depicted in Fig. 9.19,

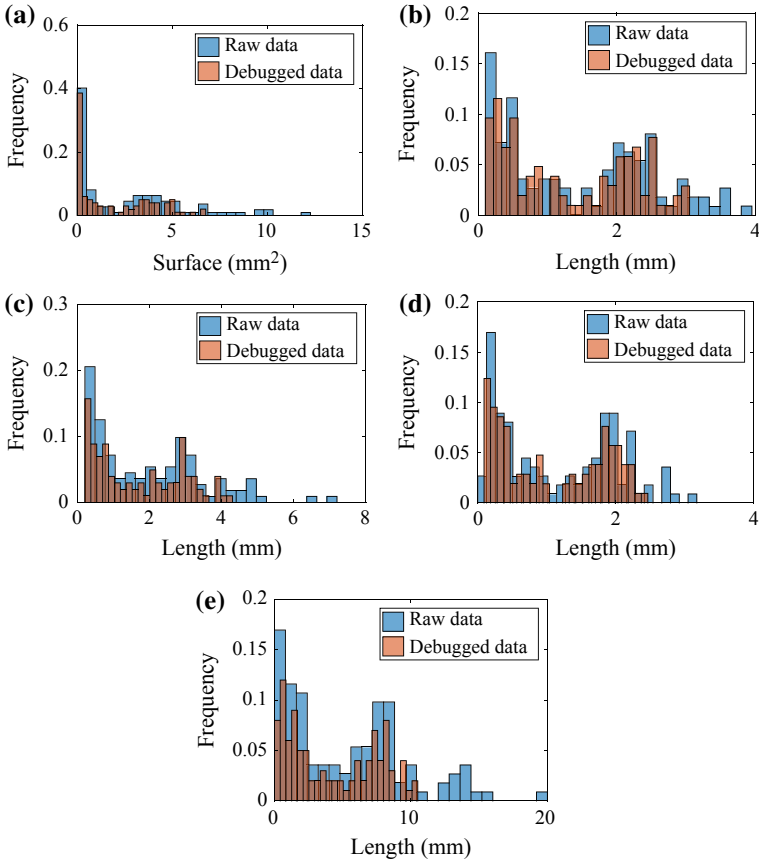


Fig. 9.18 Bubble shape histograms from experiment #2: **a** area, **b** equivalent diameter, **c** major half-axis, **d** minor half-axis, **e** perimeter. *Raw data* Data obtained without considering bubble superposition. *Debugged data* Data obtained after debugging 1.5σ times its shape parameter

Table 9.5 Statistical analysis with raw data from experiment #1 with 0.01 MPa

Parameter	Mode	Mean	Minimum	Maximum	Range	Typical deviation
Area (mm ²)	0.02	1.86	0.02	7.68	7.66	2.03
Equivalent diameter (mm)	0.15	1.20	0.15	3.13	2.97	0.97
Major half-axis (mm)	0.27	1.48	0.26	4.50	4.24	1.17
Minor half-axis (mm)	0.09	1.06	0.09	3.09	3.00	0.90
Perimeter (mm)	0.31	4.32	0.31	13.95	13.64	3.81

Table 9.6 Statistical analysis with debugged data from experiment #1 with 0.01 MPa

Parameter	Mode	Mean	Minimum	Maximum	Range	Typical deviation
Area (mm ²)	0.02	1.56	0.02	4.88	4.86	1.68
Equivalent diameter (mm)	0.15	1.11	0.15	2.53	2.38	0.90
Major half-axis (mm)	0.27	1.35	0.26	3.14	2.88	1.06
Minor half-axis (mm)	0.09	0.99	0.09	2.38	2.29	0.84
Perimeter (mm)	0.31	3.86	0.31	9.78	9.47	3.44

Table 9.7 Statistical analysis with raw data from experiment #2 with 0.05 MPa

Parameter	Mode	Mean	Minimum	Maximum	Range	Typical deviation
Area (mm ²)	0.05	2.54	0.02	12.15	12.13	2.83
Equivalent diameter (mm)	0.25	1.46	0.15	3.93	3.78	1.06
Major half-axis (mm)	0.27	1.99	0.26	7.12	6.86	1.54
Minor half-axis (mm)	0.17	1.18	0.09	3.15	3.06	0.85
Perimeter (mm)	0.31	5.30	0.31	19.77	19.46	4.38

Table 9.8 Statistical analysis with debugged data from experiment #2 with 0.05 MPa

Parameter	Mode	Mean	Minimum	Maximum	Range	Typical deviation
Area (mm ²)	0.05	1.87	0.02	6.66	6.64	1.98
Equivalent diameter (mm)	0.25	1.30	0.15	3.03	2.88	0.94
Major half-axis (mm)	0.27	1.68	0.26	4.29	4.03	1.21
Minor half-axis (mm)	0.17	1.07	0.09	2.45	2.36	0.77
Perimeter (mm)	0.31	4.24	0.31	10.43	10.12	3.26

it provides a higher number of speed intervals (histogram columns) than SIFT and SURF methods. This behaviour in Optical Flow relies on the detected areas of the image where there is no movement, which is also related with the homogeneity of the objects to detect. It should be highlighted that both the SIFT and SURF methods do not perform a pixel to pixel movement detection, instead they search the most interesting points and track them without taking into account the areas of the image without movement. For this reason, SIFT and SURF estimation methods can be used to obtain an interval of lower speeds.

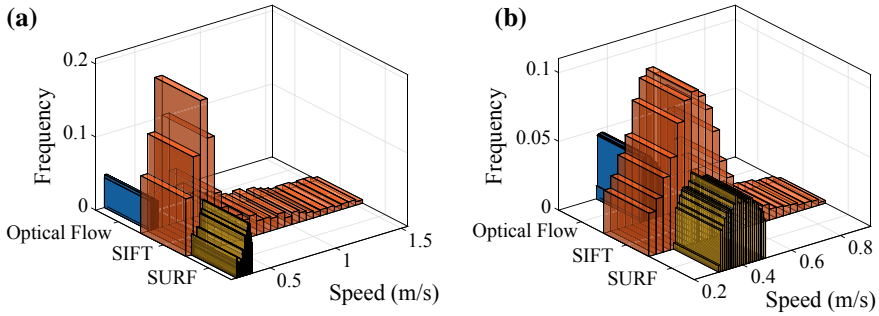


Fig. 9.19 Histograms of the bubbles’ speed estimated with SURF, SIFT and optical flow methods for: **a** experiment #1 with 0.01 MPa, and **b** experiment #2 with 0.05 MPa

Table 9.9 Bubbles’ speed statistical results for different estimation methods

Experiment	#1		#2	
Calculation method	Mean	Mode	Mean	Mode
Optical flow (m/s)	0.30	0.30	0.40	0.40
SIFT (m/s)	0.51	0.30	0.42	0.40
SURF (m/s)	0.32	0.30	0.40	0.40

9.4 Conclusion

In this chapter, we have proposed and evaluated experimentally the performance of a novel photonic sensor architecture comprising a laser diode and a CMOS array with a tunable optic lens designed to evaluate the morphology and speed of the bubbles present in a bubbly water column. With the information provided by the photonic sensor, it is possible to count the number of bubbles present in the flow and characterize them in term of size and speed getting statistical parameters. This sensor finds application in medical and industrial applications as it can be used to monitor the behaviour of a multiphase water flow. In addition, it can be used to increase efficiency and security in industrial environments (such as fabrication processes or power generation plants) while providing a longer life time not only of the structure but also of the hydraulic components, such as turbines and marine propeller screws, amongst others.

We have described the optical subsystems comprising the proposed laser-assisted CMOS sensor and the associated algorithms developed to process the data. The optical sensor is formed by a CMOS array that captures high-speed frames of the bubbles illuminated with a laser diode. A laser diode emitting at $\lambda = 532 \text{ nm}$ is selected to optimize the CMOS sensing performance and enhance the contrast of the bubbles present in the water flow.

The experimental data gathered in the laboratory demonstration with a water column indicates that bubbles’ shape depends on the diffusor porous dimensions. In

these validation examples the minor and major half-axis of the bubbles are estimated to be 0.99 and 1.35 mm, respectively for experiment #1 (void pump set to 0.01 MPa); and 1.07 and 1.68, respectively for experiment #2 (void pump set to 0.05 MPa). The mean equivalent area of the bubbles is of 1.56 mm^2 with a deviation of $\sigma = 1.68$ and, 1.87 mm^2 with a deviation of $\sigma = 1.98$. Such typical deviation appears due to the diffusor porous diameter is not uniform, which generates bubbles with different sizes emulating a real-case scenario.

The experimental results point out that a correct binarization in the post-processing of the images is necessary to calculate the air concentration in the sensed area. For this reason, the proposed laser-assisted sensor enables a higher-contrast capture of the images in order to detect the bubbles present in the flow. The post-processing algorithms are designed to obtain an adequate shape contouring of the bubbles, which enhances the processing efficiency and provide accurate results in terms of shape and velocity of bubbles. Finally, the experimental evaluation includes a comparison of three velocity estimation methods. The results point out that SIFT method is not recommended for the characterization of slow flows. SURF and Optical Flow methods can be used for evaluating velocity distributions in multiphase flows independently of their mean velocity. Despite Optical Flow method provides more velocity intervals than SURF method, both algorithms provide similar results while SURF method requires much less computation time. For example, whereas Optical Flow method required approximately 20 min to compute 500 frames at highest resolution, SURF method only required approximately 5 min. Therefore, in final implementations, the user should choose the estimation method taking into account the resolution requirements and the time they can invest to process all the information depending on their computer specifications and time available.

Acknowledgements This work was supported in part by Spain National Plan MINECO/FEDER UE RTC-2014-2232-3 HIDRASENSE and TEC2015-70858-C2-1-R XCORE projects. BIOFRAC-TIVE project with IIS La Fe is also acknowledged. Sergio L. Carrasco-Ortiz work was supported by UPV predoc FPI-UPV-2017 program. Maria Morant work was partly supported by Spain Juan de la Cierva IJCI-2016-27578 grant.

References

1. K. Iyer, S. Shridharani, S. Arunkumar, M. Venkatesan, Application of image processing for a bubble column reactor, in *Computational Intelligence and Computing Research (ICCIC)* (2013)
2. Y. Bian, F. Dong, H. Wang, Reconstruction of rising bubble with digital image processing method, in *Instrumentation and Measurement Technology Conference (I2MTC)* (2011)
3. S.L. Carrasco-Ortiz, E. Valero, M. Morant, R. Llorente, Photonic sensing and characterisation of multiphase water flows, in *PHOTOPTICS 2018—6th International Conference on Photonics, Optics and Laser Technology* (2018), pp. 55–62
4. R. Ramazani-Rend, S. Chelikani, E.M. Sparrow, J.P. Abraham, Experimental and numerical investigation of orbital atherectomy: absence of cavitation. *J. Biomed. Sci. Eng.* **3**(11), 1108–1116 (2010)

5. N. Li, L. Jia, P. Zhang, Detection and volume estimation of bubbles in blood circuit of hemodialysis by morphological image processing, in *Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)* (2015)
6. P. Jonsson, Evaluation of air micro bubbles in dialysis systems in vitro. *Am. J. Kidney Dis.* **63**(5) (2014)
7. R.F. Muddle, H.B.M. Schulte, H.E.A. Vand den Akker, Analysis of a bubbling 2-D gas-fluidized bed using image processing. *Powder Technol.* **81**(2), 149–159 (1994)
8. C.E. Brennen, *Fundamentals of Multiphase Flow* (Cambridge University Press, 2005)
9. Heliciel, s.f. Mecaflux Heliciel. <https://www.heliciel.com/en/aerodynamique-hydrodynamique/cavitation-%20helices-hydrofoils.htm>. Last accessed 27 March 2018
10. ETH Zurich, Laboratory of Hydraulics, Hydrology and Glaciology, s.f. Air detrainment of high-speed waters flows. http://www.vwa.ethz.ch/people/hy/archive/hy_151_air_detrainment. Last accessed 12 Dec 2017
11. G.V.S.S. Mittapalli, M. Madavi, R. Anirudh, E.R. Goud, Analysis of pressures on Nagarjuna Sagar spillway. *EC Agric.* **1**(1), 23–34 (2015)
12. H.C. Buckland, I. Masters, J.A.C. Orme, T. Baker, Cavitation inception and simulation in blade element momentum theory for modelling tidal stream turbine. Part A: *J. Power Energy*, Juny **227**(4), 479–485 (2013)
13. L. Shan-Yong, W. Jiang-An, Z. Si-Guang et al., Laser detection method of ship wake bubbles based on multiple scattering intensity and polarization characteristics. *Acta Phys. Sin.* **62**(6), 060704 (2013)
14. A.J. Peterka, The effect of entrained air on cavitation pitting, in *Proceedings of Minnesorta International Hydraulic Convention* (1983), pp. 507–518
15. H. Chanson, Drag reduction in open channel. *J. Hydraul. Res.* **32**(1), 87–101 (1994)
16. M. Sommerfeld, Particle dispersion in turbulent flow: the effect of particle size distribution. *Part. Part. Syst. Charact.* **7**(1–4), 209–220 (1990)
17. Y.U. Edel, Guide to protecting spillway structures from cavitation. *Power Technol. Eng. (Former. Hydrotech. Constr.)* **11**(9), 948–949 (1977)
18. P. Volkart, P. Rutschmann, Rapid flow in spillway chutes with and without deflectors. A model-prototype comparison, in *Proceedings of the Symposium on Scale Effects in Modeling Hydraulic Research* (1984), pp. 3–6
19. C. Cheng, C. Webster, J. Wong, Cavitation-resistant coatings for concrete hydraulic structures. *ACI Mater. J.* **87**(6), 594–601 (1990)
20. R. Duraiswami, S. Prabhukumar, G.L. Chahine, Bubble counting using an inverse acoustic scattering method. *J. Acoust. Soc. Am.* **104**(5), 2699–2717 (1998)
21. M. Chudina, Noise as an indicator of cavitation in a centrifugal pump. *Acoust. Phys.* **49**(4), 463–474 (2003)
22. I. Karassik, T. McGuire, *Centrifugal Pumps*, 2nd edn. (Pergamon, New York, 1997)
23. E.A. Brujan, G.S. Keen, A. Vogel, J.R. Blake, The final stage of the collapse of a cavitation bubble close to a rigid boundary. *Phys. Fluids* **14**(1), 85–92 (2002)
24. T. O’hern et al., Development of an electrical impedance tomography system for an air-water vertical water-column, in *Forum on Measurement Techniques in Multiphase Flows ASME* (1995), pp. 1–9
25. H. Chanson, Air bubble entrainment in open channels. Flow structure and bubble size distributions. *Int. J. Multiph. Flow* **23**(1), 193–203 (1997)
26. I. Bankman, *Handbook of Medical Image Processing and Analysis* (s.l.:Academic Press, 2008)
27. R. Kountchev, K. Nakamatsu, *New Approaches in Intelligent Image Analysis: Techniques, Methodologies and Applications* (Springer, Sofia, 2016)
28. J.J. Esqueda Elizondo, L.E. Palafox Maestre, s.f. Fundamentos de procesamiento de imágenes (s.l.:Universidad Autónoma de Baja California)
29. MathWorks, Inc, s.f. strel. <https://es.mathworks.com/help/images/ref/strel-class.html>. Last accessed 19 Dec 2016
30. S. Nagabhushana, *Computer Vision and Image Processing* (New Age International (P), New Delhi, 2005)

31. H. Badioze Zaman et al., *Advances in Visual Informatics* (Springer, Bangi, Malaysia, 2015)
32. A. Vedaldi, B. Fulkerson, VLFeat—an open and portable library of computer vision algorithms. <http://www.vlfeat.org/> (2008)
33. B. Herbert, T. Tinne, L. Van Gool, SURF: speeded up robust features, in *9th European Conference on Computer Vision* (2006), pp. 404–417

Chapter 10

Numerical Simulation of the Plasma Inside a Glow Discharge Millimeter Wave Detector



Cemre Kusoglu-Sarikaya, Demiral Akbar and Hakan Altan

Abstract Nowadays, Terahertz (THz) and mm-waves are encountered in many branches such as security, industry and medicine. However, limitations in existing commercial detectors in terms of cost, speed and responsivity prevent this highly advantageous region of the electromagnetic spectrum from being used more efficiently. For these reasons, neon indicator lamps, otherwise known as glow discharge detectors (GDDs), have succeeded in attracting the attention of researchers, not only because they are cheap, but also because of their better performance to other technologies. In these detectors, detection occurs as a result of the interaction of THz/mm-wave with the plasma in the lamp. Although this interaction has been tried to be explained qualitatively using various analytical models, there has been no accurate quantitative explanation about this interaction in the literature. The problem is mainly due to the confined plasma environment which can be difficult to resolve with equilibrium models. As known, GDDs are non-local thermal equilibrium plasma lamps, and such plasmas need to be modeled with a kinetic approach. For that reason, in this study, parallel 1d3v Particle in Cell/Monte Carlo Collision (PIC/MCC) simulation of the plasma in the neon lamp is performed. This simulation will form the basis for the planned investigations into understanding the effects of THz/mm-waves on the plasma. The kinetic approach employed in the simulation allows us to accurately understand and predict the plasma parameters causing the glow discharge in GDDs. The results are checked using experiments performed on home-built discharge glow chambers with similar gap dimensions.

C. Kusoglu-Sarikaya (✉) · H. Altan
Middle East Technical University, Department of Physics, 06800 Ankara, Turkey
e-mail: ckusoglu@metu.edu.tr

H. Altan
e-mail: haltan@metu.edu.tr

D. Akbar
Department of Mechanical Engineering, Ostim Technical University, 06370 Ankara, Turkey
e-mail: demiral.akbar@ostimteknik.edu.tr

10.1 Introduction

Although THz and mm-waves can be used in many areas, the lack of a cheap, fast and room-temperature detection mechanisms limits the usability of these waves. For these reasons, it is of great importance to investigate the usability of GDDs for THz and mm-wave detection since they are fast (less than 1 μ s rise time), simple, cheap and they can operate at room temperature [1]. In addition, recent studies [2, 3] have yielded promising results. When compared with the commercial detectors available today, it was observed that although glow discharge detectors have the same performance as pyroelectric detectors, Golay cells, bolometers and Schottky diodes, GDDs are better in terms of speed and response time [2, 4]. Also, the fact that GDDs are relatively cheap gives them a great advantage over all of these other expensive detectors.

Detectability of electromagnetic radiation using dc gas discharge plasmas began to be explored by researchers in the 1950s as a result of Burroughs and Bronwell's observation [5] of changes in the discharge current caused by absorption of microwave radiation. After this observation, although the detectability of microwave radiation in different glow discharge regions was investigated [6, 7] and an attempt was made to explain the detection mechanism [8, 9], a definite conclusion was not reached. While Lobov [8] argued that diffusion current could occur as a result of the absorption of microwave radiation, Severin [9] suggested that the absorbed radiation energy could cause a reduction in the ionization potential.

The usability of indicator lamps as microwave radiation detectors was firstly investigated by Farhat [10]. Meanwhile, Severin [9, 11] showed that millimeter wave frequencies can also be detected using glow discharge plasma. Later, Kopeika [12] observed in his studies that the spectral response ranges of GDDs began at microwave and extended to x-ray. GDD's detection mechanism is activated by interaction of plasma in the lamps with mm-wave or THz radiation. Since the frequency of the plasma is in the sub-GHz to GHz range, such waves can penetrate into the plasma. Taking into account the theories previously made for the mechanism of interaction between microwave and plasma [8, 9], Kopeika [13] demonstrated under what conditions enhanced diffusion and enhanced ionization occur. Although all works done so far to explain the interaction mechanism has given an idea about this mechanism, they cannot precisely explain the physical events that took place during the interaction. For this reason, in this study, the plasma in GDD was simulated using the parallel 1d3v PIC/MCC code [14]. By using this simulation, the aim here is to understand the basic plasma parameters such as eedf, iedf, ion and electron densities, the electric field and potential that cause the expected normal and abnormal glow discharge in these lamps. For a better understanding of the simulation results, a brief description of the gas discharge plasma will be given in the following section. Then, the question of why their use with THz and mm-waves attract so much attention will be tried to be answered in Sect. 10.3. In Sect. 10.4, glow discharge detectors which have many promising features in terms of detection of these radiations will be mentioned. Then, after the validation of the code in Sect. 10.5, the procedures followed for simulating

these GDDs will be described in Sect. 10.6. Finally, the simulation results will be discussed in Sects. 10.7 and 10.8.

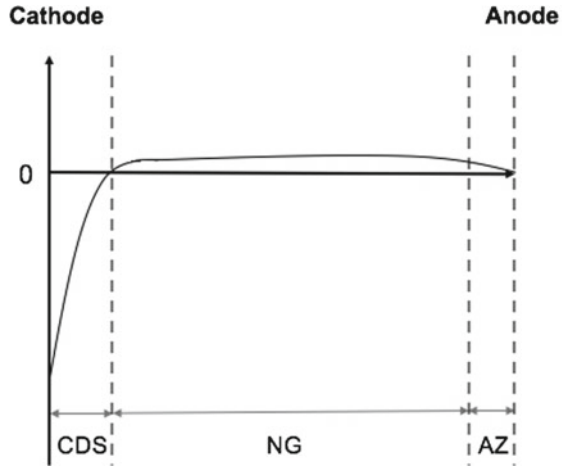
10.2 Gas Discharge Plasmas

Gas discharge plasmas occur when an electric current flows through the plasma or when the electromagnetic radiation is absorbed by the plasma. The best known example of discharge in nature is lightning. Due to their thermodynamic properties, discharges can be classified into two types: local thermal equilibrium (LTE) and non local thermal equilibrium (non-LTE) discharges. In LTE, all plasma particles are at the same temperature, while in non-LTE, which is the scope of this chapter, the temperature of electrons is much higher than that of other plasma particles. In fact, this can be explained by pressure. At high gas pressure ($10\text{--}10^3$ kPa), there are too many collisions in the plasma, resulting in efficient energy exchange. As a result, the plasma particles will have equal temperatures. On the contrary, different temperature values will be observed in the plasma particles at low gas pressure ($10^{-4}\text{--}1$ kPa) since there will be a few collisions and consequently no efficient energy exchange between the particles.

Gas discharge plasmas can also be formed in laboratories using sources such as direct current (dc), capacitively coupled radio frequency (rf), inductively coupled rf and microwaves [15, 16]. These sources vary according to the application to be used. Today, many technological applications, such as surface modification [16, 17], ozone generation [18], plasma thrusters and rocket propulsion [19], use gas discharge plasmas. Especially in light production systems such as in lamps [20] and in plasma displays [21], they are used efficiently.

In this study, we will focus on the dc glow discharges that occur in GDDs. The regions formed in such discharges are summarized in Fig. 10.1. It can be seen that although the applied potential difference between the anode and the cathode is constant in dc glow discharge, the potential and the electric field are not uniformly distributed between the electrodes. In the region next to the cathode, called the cathode dark space (CDS), the potential drops rapidly almost to zero and this drop causes a strong electric field. This region is also called the sheath region. After a sudden drop, the potential remains almost constant in the area called the negative glow (NG). Finally, the potential returns to zero in the anode zone (AZ). When the inter-electrode distance is longer, two separate regions called Faraday dark space (FDS) and positive column (PC) emerge. The negative electric field in these regions, provides the electrons move toward the anode. Since the distance between the electrodes is about 1 mm, these regions are not formed in GDDs. The lamps having an inter-electrode distance of 1–2 mm, have only one glow region close to the cathode and most likely this region can be characterized as the negative glow [22].

Fig. 10.1 The dc glow discharge regions at short inter-electrode distance and/or low pressure



10.3 THz and MM-Waves

It is known that the majority of the universe is filled with mm-wave and THz energy (0.123–12.3 meV), although it has a low enough energy to be detected [23]. This energy value falls in the range of 1.43–143 K, which is considerably lower than the earth's ambient background temperature. This type of energy is mostly emitted from cold dust in the interstellar environment [24]. MM-waves, are in the range of 30–300 GHz in the electromagnetic spectrum, right next to the THz waves (300 GHz–3 THz). As the name implies, the wavelengths of these waves are in millimeter order. The presence of resonant absorption of THz/mm-waves in water and oxygen molecules limits their usability in long distance communications [25]. However, they can be used in many applications at distances of about 10m [26]. In addition, their short wavelengths allow to transmit a lot of data at one time.

Since THz and mm-waves have a wide variety of properties, it is possible to use them in many areas. First of all, these waves are very beneficial in the field of security [27, 28] because they can penetrate many non-conductive materials and high resolution images can be obtained due to their short wavelengths. In addition, the presence of characteristic spectra of some explosives at the THz frequency gives them a great advantage in terms of detection as well as identification. As known, THz/mm waves are non-ionizing radiations and can be used on humans. With this feature, it has been found that these radiations can also be used in pharmaceutical [29, 30], medical [31, 32] and non-destructive testing areas [33].

Despite all these advantageous features, the available sources and detectors for THz/mm-waves are quite limited. In particular, the fact that existing commercial detectors are very expensive restricts the efficient use of these waves. For this reason, in this study, GDDs with a cheap and promising THz/mm-wave detection mechanism is tried to be understood by examining the parameters of the plasma formed inside them.

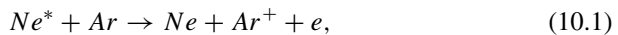
10.4 Glow Discharge Detectors

Some of the more common, currently available commercial detectors for THz/mm-waves are Schottky diodes, Golay cells, pyroelectric detectors and bolometers [23]. The investigations [2–4] have shown that GDDs are more advantageous than these detectors because they are cheap, fast and operate at room temperature. They also show relatively low sensitivity to changes in ambient temperature [9, 11, 35].

GDDs are glass tubes containing two electrodes and a low-pressure noble gas, some have a phosphorescent coating inside the glass. A sufficient potential difference between the electrodes causes the ionization of the gas and thus the formation of electrons and ions. As the negatively charged particles move toward the anode, the positively charged particles move toward the cathode, causing secondary electron emission in the cathode, which plays an important role in the sustainability of the generated plasma. The light is the result of excitation of the gas atoms by the electrons in the plasma. The excited gas atoms fall to a lower energy level by emitting photons in a color that is characteristic of the gas.

Different noble gases can be used in GDDs. Since each gas has different characteristics such as molecular weight and ionization energy, the plasma parameters of GDDs using different noble gases are different. This difference affects the THz/mm-wave detectability [38]. The lowest efficiency from GDD is obtained when helium gas is used. Since the ionization energy of helium is the highest among the noble gases, the discharge potential of this gas is also the highest. As a result, the electric field will be higher than the THz/mm-wave electric field, so the radiation effect will not be observed efficiently. Although argon, krypton, and xenon gases have lower ionization energies, they require a higher discharge potential to reach the cathode and form secondary electrons, since they are heavier. This, as mentioned for the helium gas, reduces the efficiency of the detector. Neon gas with low atomic weight has the lowest discharge potential, which leads to the ability to detect THz/mm-waves with the greatest responsivity [38].

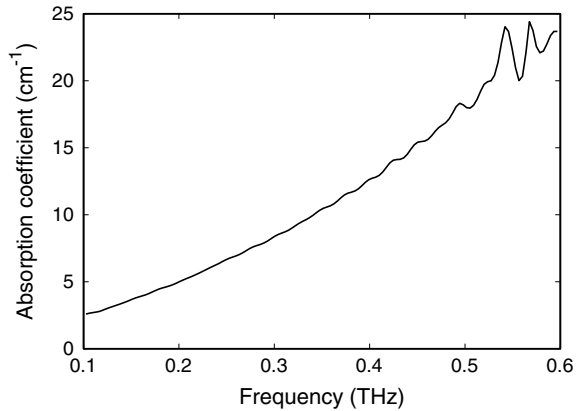
In GDDs, noble gases can be used in mixture form as well as in pure form. Generally, Ne-Ar mixture is used to reduce the discharge potential of the neon gas to lower values. Since the excited neon atoms have higher energy than the ionization energy of argon gas, excited neon atoms can ionize the argon gas. The reaction, which has the form



is called Penning ionization. With this reaction, ionization can occur at lower voltage values [39]. In Ne-Ar mixture, when 1 or less percent of argon gas is used, Penning ionization occurs at a significant level. However, when the percentage of argon gas is further increased, the argon gas will begin to be ionized directly by the electrons, thus the effect of Penning ionization will be reduced [40].

In GDDs, as well as the gas used, the operation in different dc glow discharge regions affects THz/mm-wave detection. In the Townsend region, a large amount of photons are needed for THz/mm-wave detection, since electrons mostly have low

Fig. 10.2 Absorption coefficient of the glass wall typically used in a commercial GDD



energy. In the subnormal glow region, discharge current is very low and mainly caused by ion-induced secondary electron emission in the cathode. For that reason, the main contribution to the noise comes from this secondary emission process [41, 42]. It is not appropriate to work in this region because this noise will cover the THz/mm-wave signal [25]. In the arc discharge region, it is very difficult for GDDs to work stably since large amounts of thermal noise will occur. GDDs need to be operated in normal or abnormal glow regions in order to detect THz/mm-waves efficiently. Because, in these regions, the discharge current is the result of both ion-induced secondary electron emissions and electron-neutral collisions. As a result of the interaction of THz/mm-waves with the electrons, the increase in energy of the electrons will affect electron-neutral collisions and therefore cause a change in the discharge current.

Since the frequency of the THz/mm-waves is greater than the plasma frequency (in the sub-GHz to GHz range [1, 43]), these waves can pass through the plasma in neon lamps. It is thought that this interaction results in a change in the discharge current through cascade ionization and diffusion current [13]. Cascade ionization is the process by which electrons ionize neutrals in a shorter time with the energy they absorb from electromagnetic waves. This event results in an increase in the discharge current. Diffusion current is, however, the fact that electrons diffuse into the electrodes without performing ionization with the energy they absorb, and this diffusion causes a decrease in the discharge current. Both of these events are thought to occur during the interaction with plasma in the GDDs, but one dominates [13]. Indeed, in some studies with GDD, it was observed that the discharge current increased after interaction [1, 13], whereas in the other studies it was observed that the discharge current had decreased [2]. Although these studies have made a great contribution to the understanding of GDDs, they have been limited to analytical models and qualitative explanations. It is still unclear what kind of physical events take place during the interaction of THz/mm-wave and plasma.

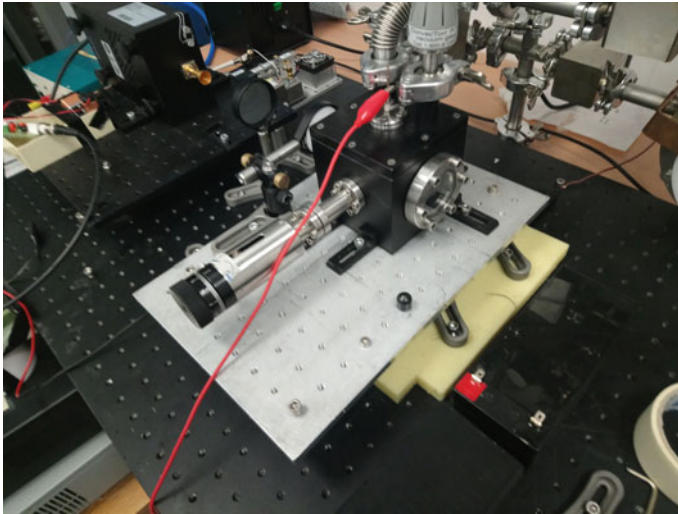


Fig. 10.3 Plasma discharge chamber. The chamber can be filled by a single or mixture of gases. The distance between the anode and cathode can be adjusted with $10\ \mu\text{m}$ precision

In addition, since neon lamps are not designed as detectors, they have to be optimized. For example, in Fig. 10.2, the absorption coefficient of the glass wall of the GDDs is plotted according to the frequency. As can be seen, there is a significant loss in power of the THz wave after the glass wall. First of all, a different material must be used in order to prevent this loss. Furthermore, since the distance between the electrodes affects the plasma, determining the optimum value of this distance is of great importance. For this reason, the plasma discharge chamber shown in Fig. 10.3 has been developed in our laboratory to determine the optimum parameters required for the use of GDDs as detectors. The dimension of this chamber is $10 \times 10 \times 10\ \text{cm}^3$. It has two quartz windows with a diameter of 40 mm which allow the transmission of THz radiation. The inter-electrode distance can be increased to 2 cm with $10\ \mu\text{m}$ resolution. With two feedthroughs on the top of the chamber, instantaneous pressure measurement and distribution of gas mixtures can be made. Up to three different gases can be mixed with the gas mixer system. Also, the rotary vane pump allows the pressure in the chamber to be reduced to 10^{-2} torr. To be able to use this chamber to find the optimum parameters efficiently, it is first necessary to understand the plasma parameters in GDDs and the effects of these parameters on the THz/mm-wave and plasma interaction. For this purpose, in this study, the parallel 1d3v PIC/MCC code [14] is used to compute the basic plasma parameters such as eedf, iedf, ion and electron densities, the electric field and potential inside the GDD.

10.5 Validation of the PIC/MCC Simulation Code

As known, GDDs are low pressure non-LTE plasma lamps. For this reason, as mentioned in Sect. 10.2, since the energy exchange is not done efficiently in these type of plasmas, the plasma particles do not have the same temperature. In order to understand the behavior of such plasma particles adequately, a kinetic approach is required. This approach can be obtained by using the distribution function $f(\mathbf{r}, \mathbf{v}, t)$ defined by the Boltzmann equation,

$$\left[\frac{\partial}{\partial t} + \mathbf{v} \nabla_{\mathbf{r}} + \mathbf{a} \nabla_{\mathbf{v}} \right] f(\mathbf{r}, \mathbf{v}, t) = \left(\frac{\partial f}{\partial t} \right)_{\text{coll.}}, \quad (10.2)$$

where, \mathbf{v} , \mathbf{r} and \mathbf{a} are the velocity, position and acceleration of the plasma particles, respectively. However, since the direct solution of this equation is highly complicated mathematically, particle simulation methods that give the same results [44, 45] can be used to explain the kinetic behavior of particles. One of these methods is the well-known PIC/MCC method. In this study, the plasma in the GDD is simulated using the parallel 1d3v PIC/MCC code [14]. The code is written in Fortran 90 language and parallelized by using Open MPI library for performance enhancement. However, since the gas pressure and the gas mixture ratios (if gas mixture is used) used in neon indicator lamp are not known exactly, before simulating GDDs, an experiment was performed with argon gas in the home built chamber with certain parameters and the discharge current obtained from the experiment is compared with that obtained from the simulation. Inside the chamber, the argon gas pressure is held constant at 1.7 torr and the background pressure is reduced to about 0.01 torr. The distance between the electrodes is set to be 1.2 mm.

As can be seen from Fig. 10.4, under the specified conditions, when 350 V is applied to the argon gas, a discharge current of 0.38 ± 0.04 mA is obtained. Taking this result into consideration, the experiment is simulated. The types of collisions

Fig. 10.4 The voltage-current characteristic obtained for argon gas in the plasma discharge chamber

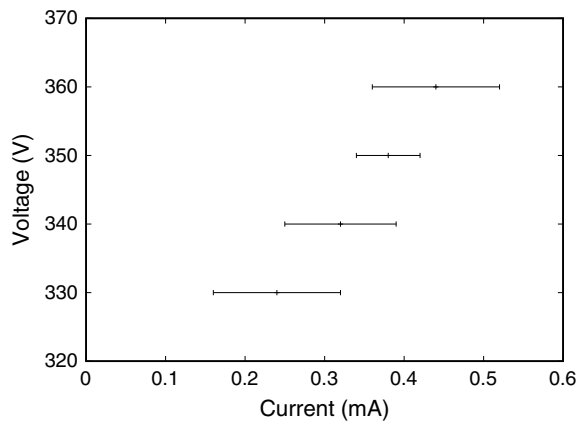


Fig. 10.5 Elastic, excitation, ionization cross-section curves for electrons in argon gas. These data were taken from [46]

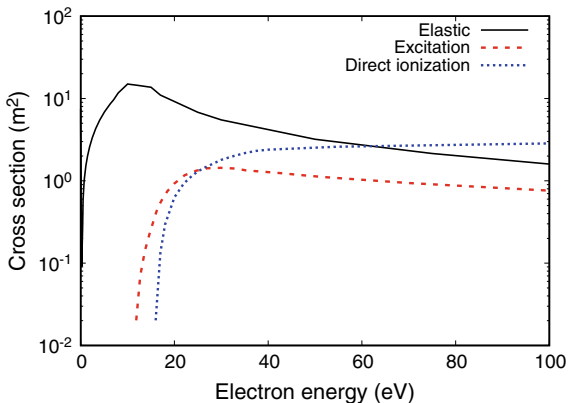


Fig. 10.6 Isotropic and backward scattering cross-section curves for ions in argon gas. These data were taken from [47]

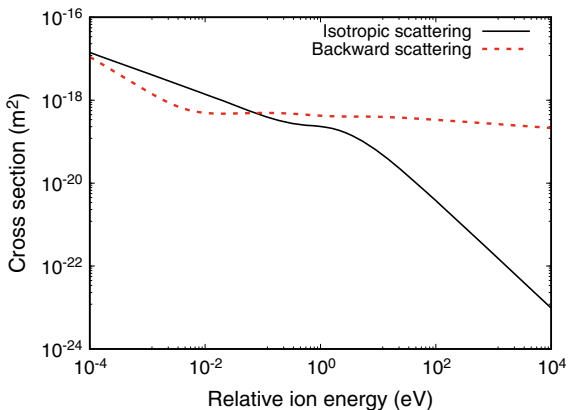


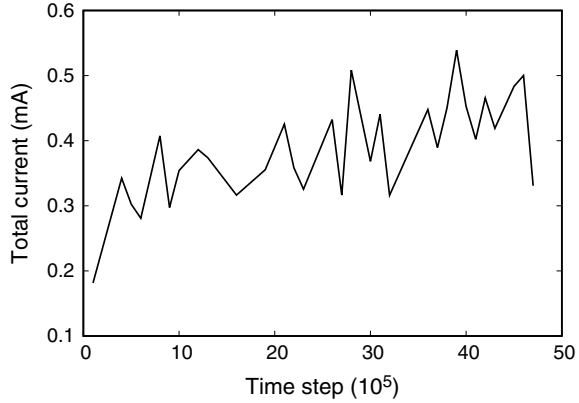
Table 10.1 The numerical parameters used in the PIC/MCC simulation of argon gas

Weighting (10^8)	2.4
Grid number	600
Time step (10^{-12} s)	4
Gas temperature (K)	300
Secondary electron emission coefficient (γ)	0.13
Reflection coefficient (r)	0.2

considered in the simulation, and the cross-section curves of these collisions [46, 47] are shown in Figs. 10.5 and 10.6. Numerical parameters used are summarized in Table 10.1.

As a result of the simulation, the total current, which contains the electron, ion and displacement current, is obtained over time (Fig. 10.7). It is apparent that the total current starts to converge around 0.4 mA after the 4×10^6 th time step, which corre-

Fig. 10.7 The change of the total current with time obtained in the simulation of Ar glow discharge. Each time step corresponds to 4×10^{-12} s



sponds to the real time of $16 \mu\text{s}$. This converged result is approximately in agreement with the discharge current obtained from the experiment. This compatibility supports the usability of the code for simulating a commercially available GDD.

10.6 Parallel 1d3v PIC/MCC Simulation of GDD

By using parallel 1d3v PIC/MCC code, GDD is simulated using neon gas and Ne-Ar mixture, separately. Between electrons and neutrals, elastic, excitation and ionization reactions are considered in the code. The cross section curves describing these reactions are shown in Fig. 10.8.

As shown in Fig. 10.8, only excitations to $Ne(^3P_2)$ are considered for the excitation reaction. This is because the metastable atoms at this level are known to be responsible for the Penning ionization of the argon gas [1]. In addition, in order to

Fig. 10.8 Elastic, excitation, ionization cross-section curves for electrons in neon gas. These data was taken from [48]

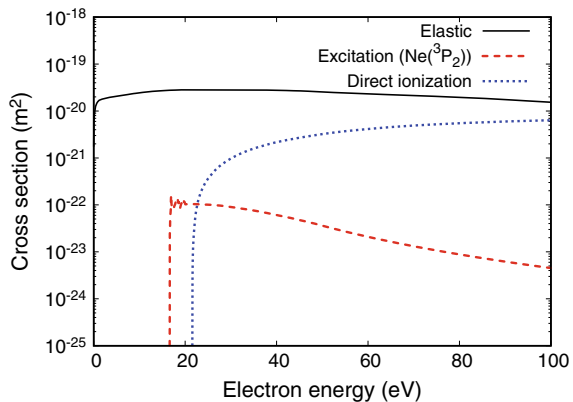


Table 10.2 The physical and numerical parameters used in the GDD simulation

Weighting (10^8)	2.6
Grid number	600
Time step (10^{-12} s)	4
Distance (mm)	1.3
Gas temperature (K)	300
Pressure (torr)	25
Voltage (volt)	115

prevent excessive time consumption in the simulation, the density distribution of excited atoms ($Ne(^3P_2)$) is investigated by using the fluid approximation,

$$\frac{\partial n}{\partial t} - D \frac{\partial^2 n}{\partial x^2} = S, \quad (10.3)$$

where D is the diffusion coefficient taken [49, 50] as $150 \text{ cm}^2\text{s}^{-1}$ torr and S is the excitation source term. This approximation is combined with the PIC/MCC model.

Between ions and neutrals, it is assumed that there are basically two separate reactions: isotropic scattering and charge transfer. Cross-section values for these reactions were taken as 2×10^{-19} and $3 \times 10^{-19} \text{ m}^2$ [51], respectively.

It is well known that the ion-induced secondary electron emission, which occurs in the cathode, plays a major role in the sustainability of the plasma in discharge lamps. This reaction is defined by a secondary electron emission coefficient, γ . In our previous work [52], this coefficient was taken as 0.2, considering that the electrode in the neon lamp is tungsten [53]. However, it is known that coatings on electrodes are used to reduce the work function value in such lamps. In the literature, it is stated that the reduction of around fifty percent is achieved by coatings [54, 55]. For this reason, it would be more reasonable to increase the secondary electron emission coefficient by 50% to 0.3. Other parameters used in the simulation are summarized in Table 10.2.

10.7 Simulation Results (Pure Neon Gas)

Figure 10.9 shows the potential and electric field curves obtained from the simulation by using pure neon gas. As can be seen, the sheath region, where the electric field is strong, is seen between 1 and 1.3 mm. It is also evident that the negative glow region, where the potential is almost constant, covers a very large distance of 0.1–1 mm and no positive column or Faraday dark space regions are observed as expected.

Figure 10.10 shows the distribution of the density of electrons, ions, and excited neon atoms between the electrodes. As can be seen, the maximum density is observed

Fig. 10.9 Electric field and electric potential profiles obtained by using pure neon gas

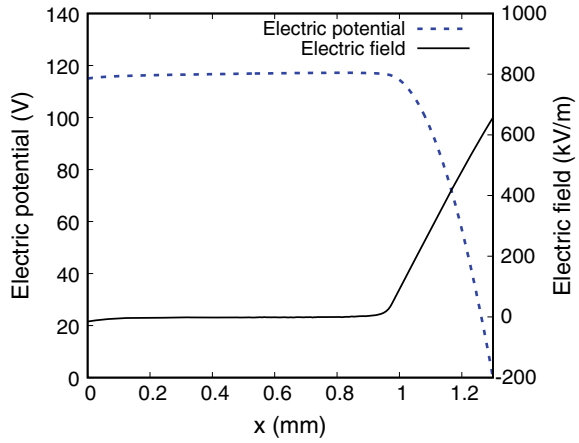


Fig. 10.10 Electron, ion and metastable neon density distribution obtained by using pure neon gas

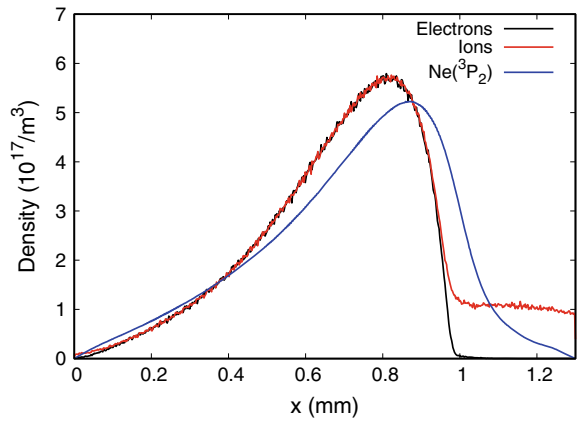
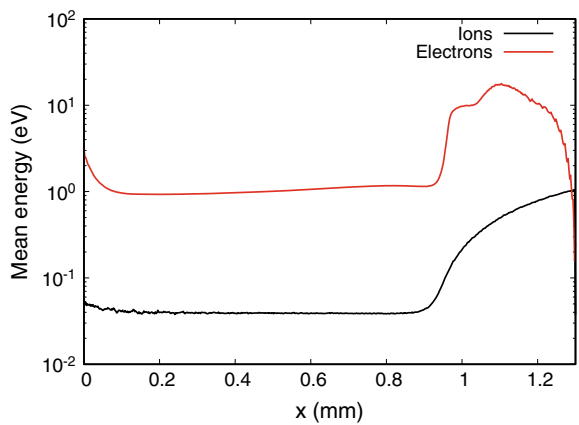


Fig. 10.11 Electron and ion mean energy distribution obtained by using pure neon gas



in a region close to the cathode. This value corresponds to the maximum plasma frequency (f_p) of 7 GHz. This supports the statement in the literature that the plasma frequency in the neon indicator lamps is on the order of a few GHz [1, 43].

The mean energy distributions of electrons and ions are shown in Fig. 10.11. As mentioned in Sect. 10.2, since GDDs are non-LTE plasma lamps, the temperature of electrons is much higher than the temperature of other plasma particles. This is evident in the graph of mean energy distribution. In addition, it is observed that particles are energized by the effect of the strong electric field in the sheath region, as expected. Low energetic particles appear dominantly in the negative glow region where quasi-neutrality is observed. This demonstrates that the particles effectively lose their energy in this region as a result of elastic collisions. Two distinct Maxwellian regions formed by these high and low energetic particles can also be seen in the distribution function profiles (Figs. 10.13 and 10.14). In the region formed by high-energy electrons, a wide peak is observed around 14 eV. By looking at the mean energy graph (Fig. 10.11), it can be seen that the electrons having these energies are

Fig. 10.12 Comparisons of electron densities obtained by using pure Ne and Ne-Ar mixture, separately

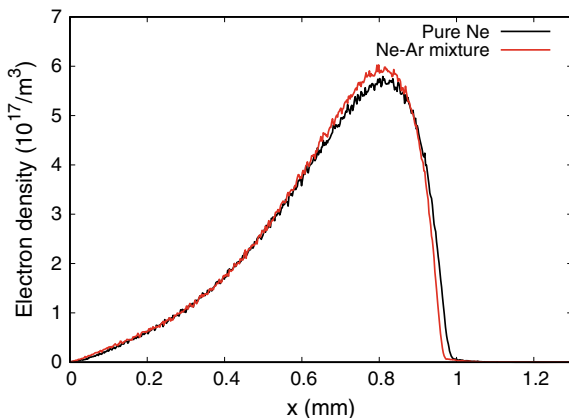


Fig. 10.13 Electron energy distribution function profiles obtained by using pure Ne and Ne-Ar mixture, separately

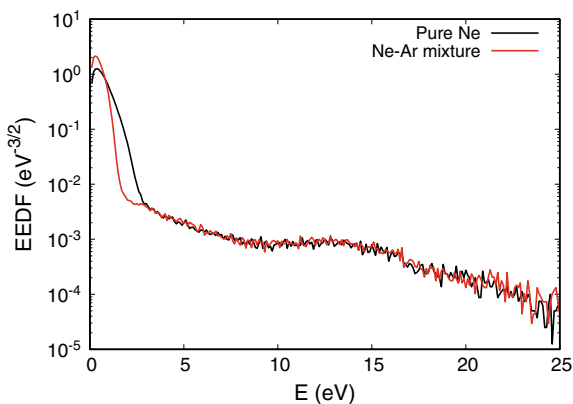
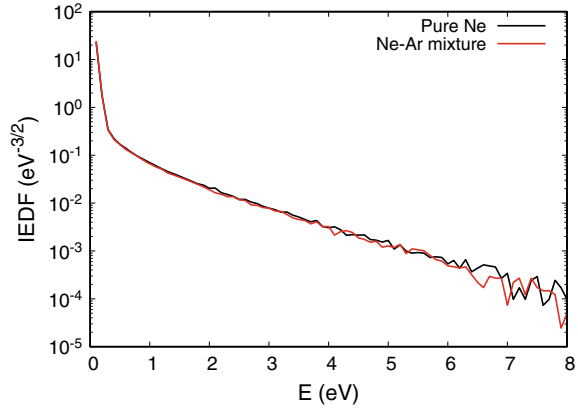


Fig. 10.14 Ion energy distribution function profiles obtained by using pure Ne and Ne-Ar mixture, separately



in a region close to the cathode. In Sect. 10.5, it was stated that the coatings on the electrodes are used to increase the ion-induced secondary electron emission efficiency in the discharge lamps. This ensures that the secondary electron emission is an effective reaction in this type of lamps. For this reason, it is thought that secondary electrons emitted from the cathode accelerate in the strong electric field and form this peak.

10.8 Simulation Results (Ne-Ar Mixture)

The simulation was repeated with the same parameters, adding only 1% argon gas to the neon gas. The electron density graph obtained from this simulation is compared with that obtained when pure neon gas is used (Fig. 10.12). As can be seen, the effect of Penning ionization, which occurs in the Ne-Ar mixture, is confirmed by an increase in density. This increase means that the profile obtained when the mixture is used can be reached with a lower potential difference than when pure neon gas is used. This explains why the gas mixture is preferred in commercial neon indicator lamps. However, since the increase observed here is negligible, it is understood that the percentage of argon should be kept below 1%.

Furthermore, the effect of gas mixture on the energy distribution profiles of the particles are also investigated (Figs. 10.13 and 10.14). While there is no change in the energy distribution of ions, the effect of low energetic electrons emitted by Penning ionization is clearly visible in the energy distribution of electrons at low energies.

10.9 Conclusions

Gas Discharge Detectors (GDDs) offer many advantages over other methods in detection of mm wave and terahertz radiation. While efforts have been made to qualitatively explain the detection mechanism behind commercially available GDDs, the low pressure non-LTE medium inside these lamps can only be truly understood quantitatively using a kinetic approach. Parallel 1d3v PIC/MCC simulation is performed to simulate this plasma medium. Since parameters such as pressure and gas mixture ratio are not known exactly for neon indicator lamp, first, an experiment with a certain pressure and voltage value was performed with a home-built plasma discharge chamber developed in our laboratory. The parallel 1d3v PIC/MCC code is found to be successful in validating the experimental measurements performed when pure argon gas is used as the discharge medium inside this chamber for similar discharge dimensions as in GDDs.

After the validation process, the GDD is simulated by using Ne and Ne-Ar mixture, respectively. First of all, it was observed from the potential and the electric field profile that the negative glow region in GDDs covers most of the inter-electrode distance. Since the distance between the electrodes is approximately 1 mm, positive column and Faraday dark space regions are not observed. Electron and ion density distributions in the negative glow region show that the plasma frequency, f_p , is in the GHz range, as indicated in the literature. Furthermore, the observed differences in the mean energy of the plasma species are in accordance with the characteristics of the non-LTE plasmas. Finally, the increase in density obtained when Ne-Ar mixture is used shows that the use of this mixture causes a decrease in the breakdown voltage. This explains why gas mixtures are preferred in commercial GDDs. However, it has been observed that the Penning effect is negligible for the percentage of argon gas used in this simulation.

In conclusion, it has been shown that the PIC/MCC simulation is useful for quantitatively determining the plasma properties of GDDs. This tool will further allow one to investigate the plasma-THz/mm-wave interaction where it is thought that the radiation increases the collisions inside the plasma and where an increase in detection is observed when the GDD is operated in the abnormal glow regime. Quantizing this interaction will allow one to better design and optimize the detection inside the glow discharge detector.

Acknowledgements The work was supported by the Scientific and Technical Research Council of Turkey (TUBITAK) 115F226. This research is also sponsored in part by the NATO Science for Peace and Security Programme under grant MD.SFPP 984775. The simulations were performed using High Performance and Grid Computing Center (TRUBA Resources) at TUBITAK ULAKBIM.

References

1. N.S. Kopeika, Glow discharge detection of long wavelength electromagnetic radiation: cascade ionization process internal signal gain and temporal and spectral response properties. *IEEE Trans. Plasma Sci.* **6**, 139–157 (1978)
2. A. Abramovich, N.S. Kopeika, D. Rozban, E. Farber, Inexpensive detector for terahertz imaging. *Appl. Opt.* **46**, 7207–7211 (2007)
3. A. Abramovich, N.S. Kopeika, D. Rozban, E. Farber, Terahertz detection mechanism of inexpensive sensitive glow discharge detector. *J. Appl. Phys.* **103**, 093306 (2008)
4. L. Hou, W. Shi, Fast terahertz continuous-wave detector based on weakly ionized plasma. *IEEE Electron Device Lett.* **33**, 1583–1585 (2012)
5. C. Burroughs, A. Bronwell, *Teleteknik* **II**, 62 (1952)
6. M.A. Lampert, A.D. White, Microwave techniques for studying discharges in gases. *Electron. Commun.* **30**, 124–128 (1953)
7. B.J. Udelson, Effect of microwave signals incident upon different regions of a dc hydrogen glow discharge. *J. Appl. Phys.* **28**, 380–381 (1957)
8. G.D. Lobov, Gas discharge detector of microwave oscillations. *Radiotekh. Electron.* **5**, 152–165 (1960)
9. P.J.W. Severin, *The Interaction of Microwaves with the Cathode Fall and Negative Glow in a Glow Discharge* (Philips Research Laboratories, Eindhoven, Netherlands, 1965)
10. N.H. Farhat, A plasma microwave power density detector. *Proc. IEEE* **52**, 1053–1054 (1964)
11. P.J.W. Severin, A.G. Van Nie, A Simple and rugged wide-band gas discharge detector for millimeter waves. *IEEE Trans. Microw. Theory Tech.* **14**, 431–436 (1966)
12. N.S. Kopeika, Theory of a fast, sensitive, submillimeter wave glow discharge detector. *Int. J. Infrared Millim. Waves* **5**, 1333–1348 (1984)
13. N.S. Kopeika, On the mechanism of glow discharge detection of microwave and millimeter-wave radiation. *Proc. IEEE* **63**, 981–982 (1975)
14. C. Kusoglu-Sarikaya, I. Rafatov, A.A. Kudryavtsev, Particle in cell/Monte Carlo collision analysis of the problem of identification of impurities in the gas by the plasma electron spectroscopy method. *Phys. Plasmas* **23**, 063524 (2016)
15. A. Bogaerts, E. Neyts, R. Gijbels, J.V.D. Mullen, Gas discharge plasmas and their applications. *Spectrochim. Acta B* **57**, 609–658 (2002)
16. M.A. Lieberman, A.J. Lichtenberg, *Principles of Plasma Discharges and Materials Processing* (Wiley, New York, 1994)
17. A. Grill, *Cold Plasma in Materials Fabrication: From Fundamentals to Applications* (IEEE Press, New York, 1994)
18. T. Hammer, Applications of plasma technology in environmental techniques. *Contrib. Plasma Phys.* **39**, 441–462 (1999)
19. D.M. Goebel, I. Katz, *Fundamental of Electric Propulsion: Ion and Hall Thrusters* (Wiley, New Jersey, 2008)
20. J.R. Coatan, A.M. Marsden, *Lamps and Lightning* (Arnold, London, 1997)
21. A. Sobel, Plasma displays. *IEEE Trans. Plasma Sci.* **19**, 1032–1047 (1991)
22. N.S. Kopeika, J. Rosenbaum, R. Kastner, Abnormal glow discharge detection of visible radiation. *Appl. Opt.* **15**, 1610–1615 (1976)
23. P.H. Siegel, Terahertz technology. *IEEE Trans. Microw. Theory Tech.* **50**, 910–928 (2002)
24. D.T. Leisawitz, W.C. Danchi, M.J. DiPirro, L.D. Feinberg, D.Y. Gezari, M. Hagopian, W.D. Langer, J.C. Mather, S.H. Moseley, M. Shao, R.F. Silverberg, J.G. Staguin, M.R. Swain, H.W. Yorke, X. Zhang, Scientific motivation and technology requirements for the SPIRIT and SPECS far-infrared/submillimeter space interferometers. *Proc. SPIE* **4013**, 36–46 (2000)
25. L. Hou, H. Park, X. Zhang, Terahertz wave imaging system based on glow discharge detector. *IEEE J. Sel. Top. Quant.* **17**, 177–182 (2011)
26. T.S. Hartwick, D.T. Hodges, D.H. Barker, F.B. Foote, Far infrared imagery. *Appl. Opt.* **15**, 1919–1922 (1976)

27. M.C. Kemp, P.F. Taday, B.E. Cole, J.A. Cluff, A.J. Fitzgerald, W.R. Tribe, Security applications of terahertz technology. *Proc. SPIE* **5070**, 44–52 (2003)
28. W.R. Tribe, D.A. Newnham, P.F. Taday, M.C. Kemp, Hidden object detection: security applications of terahertz technology. *Proc. SPIE* **5354**, 168–176 (2004)
29. P.F. Taday, Applications of terahertz spectroscopy to pharmaceutical sciences. *Philos. Trans. R. Soc. London Ser. A* **362**, 351–364 (2004)
30. C.F. Strachan, P.F. Taday, D.A. Newnham, K.C. Gordon, J.A. Zeitler, M. Pepper, T. Rades, Using terahertz pulsed spectroscopy to quantify pharmaceutical polymorphism and crystallinity. *J. Pharm. Sci.* **94**, 837–846 (2005)
31. E. Pickwell, B.E. Cole, A.J. Fitzgerald, M. Pepper, V.P. Wallace, In vivo study of human skin using pulsed terahertz radiation. *Phys. Med. Biol.* **49**, 1595–1607 (2004)
32. V.P. Wallace, A.J. Fitzgerald, S. Shankar, N. Flanagan, R. Pye, J. Cluff, D.D. Arnone, Terahertz pulsed imaging of basal cell carcinoma ex vivo and in vivo. *Br. J. Dermatol.* **151**, 424–432 (2004)
33. D.M. Mittleman, R.H. Jacobsen, M.C. Nuss, T-Ray Imaging. *IEEE J. Sel. Top. Quantum Electron.* **2**, 679–692 (1996)
34. N. Kukutsu, Y. Kado, Overview of millimeter and terahertz wave application research. *NTT Tech. Rev.* **7**, 1–6 (2009)
35. F.A. Benson, G. Mayo, Effects of ambient-temperature variations on glow-discharge tube characteristics. *J. Sci. Instrum.* **31**, 118–120 (1954)
36. Private communication. <https://www.intl-lighttech.com/specialty-light-sources/neon-lamps>. Cited 13 April 2018
37. W.G. Miller, *Using and Understanding Miniature Neon Lamps* (Howard W. Sams & Co., Inc, Indianapolis, 1969)
38. L. Hou, W. Shi, S. Chen, Z. Yan, Terahertz continuous wave detection using weakly ionized plasma in inert gases. *IEEE Electron Device Lett.* **34**, 689–691 (2013)
39. Y.P. Raizer, *Gas Discharge Physics* (Springer, Berlin, Germany, 1991)
40. E.W. McDaniel, *Collision Phenomena in Ionized Gases* (Wiley, New York, NY, USA, 1964)
41. N. S. Kopeika, J. Rosenbaum, Subnormal glow discharge detection of optical and microwave radiation. *IEEE Trans. Plasma Sci.* **4**, 51–61 (1976)
42. N.S. Kopeika, Noise spectra of commercial indicator-lamp glow-discharge detectors. *Int. J. Electron.* **39**, 209–218 (1975)
43. N.S. Kopeika, N.H. Farhat, Video detection of millimeter waves with glow discharge tubes: part I-physical description; part II-experimental results. *IEEE Trans. Electron Devices* **22**, 534–548 (1975)
44. K. Nanbu, Probability theory of electron-molecule, ion-molecule, molecule-molecule, and coulomb collisions for particle modelling of materials processing plasmas and cases. *IEEE Trans. Plasma Sci.* **28**, 971–990 (2000)
45. S. Longo, Monte carlo models of electron and ion transport in non-equilibrium plasmas. *Plasma Sources Sci. Technol.* **9**, 468–476 (2000)
46. Compilation of electron cross sections used by A. V. Phelps. http://jilawwww.colorado.edu/~avp/collision_data/electronneutral/ELECTRON.TXT. Cited 4 May 2018
47. Phelps database. <http://www.lxcat.net>. Cited 4 May 2018
48. Biagi database (Magboltz versions 8.9 and higher). <http://www.lxcat.net>. Cited 11 April 2018
49. A.V. Phelps, Diffusion, de-excitation, and three-body collision coefficients for excited neon atoms. *Phys. Rev.* **114**, 1011–1025 (1959)
50. A.V. Phelps, J.P. Molnar, Lifetimes of metastable states of noble gases. *Phys. Rev.* **89**, 1202–1212 (1953)
51. W.H. Cramer, Elastic and inelastic scattering of lowvelocity ions: He+ in Ne, Ne+ in He, and Ne+ in Ne. *J. Chem. Phys.* **28**, 688–690 (1958)
52. C. Kusoglu-Sarikaya, H. Altan, D. Akbar, Parallel 1d3v particle in cell/monte carlo collision (PIC/MCC) simulation of a glow discharge millimeter wave detector, in *Proceedings of the 6th International Conference on Photonics, Optics and Laser Technology*, vol. 1 (2018) pp. 110–115

53. H.D. Hagstrum, Auger ejection of electrons from tungsten by noble gas ions. *Phys. Rev.* **104**, 317–318 (1956)
54. H. Bruining, *Physics and Applications of Secondary Electron Emission* (Philips Research Laboratories, Netherlands, Eindhoven, 1954)
55. L.B. Loeb, *Basic Processes of Gaseous Electronics* (University of California Press, California, 1960)

Chapter 11

Dictionary Construction Method for Hyperspectral Remote Sensing Correlation Imaging



Qi Wang, Lingling Ma, Hong Xu, Yongsheng Zhou, Chuanrong Li,
Lingli Tang and Xinhong Wang

Abstract The correlation imaging technique is a novel imaging strategy which acquires the object image by the correlation reconstruction algorithm from the separated signal light field and reference light field, with the advantages such as super-resolution, anti-interference and high security. The hyperspectral remote sensing correlation imaging technique combined the correlation imaging and hyperspectral remote sensing has the ability to detect the spectral properties of ground objects more than the above advantages. The spatial and spectral images can be reconstructed from very few measurements acquired by hyperspectral correlation imaging systems via sparsity constraint, making it an effective approach to solve the process and transition problem in high spatial and spectral resolution. While in the hyperspectral remote sensing correlation imaging, due to the complexity and variance of the spatial and spectral properties of the target scene and lack of prior knowledge, it is difficult to construct an effective dictionary for the hyperspectral reconstruction. The fixed dictionaries such as DCT (Discrete Cosine Transform) dictionary and wavelet dictionary are mainly used in the reconstruction up to present, and these dictionaries contain fixed and limited characteristics, which is hard to present different hyperspectral scenes efficiently. This paper aims at the problem that in the hyperspectral remote sensing correlation imaging system the sparse representation of complex ground objects is difficult in image reconstruction, resulting in low quality of reconstructed images. By combining the sparse coding and dictionary learning theory of signal processing, a related research on the construction method of hyperspectral remote sensing sparse dictionaries is carried out. Through the construction of hyperspectral remote sensing sparse dictionaries, optimization in reconstruction, and application research in actual imaging systems, a set of sparse dictionary construction and usage methods in hyperspectral correlation imaging has been formed. Comparing with

Q. Wang · L. Ma (✉) · H. Xu · Y. Zhou · C. Li · L. Tang · X. Wang
Key Laboratory of Quantitative Remote Sensing Information Technology, Academy of
Opto-Electronics, Chinese Academy of Sciences, Beijing 100094, China
e-mail: llma@aoe.ac.cn

© Springer Nature Switzerland AG 2019
P. Ribeiro and M. Raposo (eds.), *Optics, Photonics and Laser
Technology 2018*, Springer Series in Optical Sciences 223,
https://doi.org/10.1007/978-3-030-30113-2_11

current methods such as the total variation constraints and the hyperspectral image kernel norm constraints, the hyperspectral remote sensing scene sparsity and hyperspectral image reconstruction quality have been effectively improved by the proposed method, which has guiding significance for the development of the correlation imaging field.

11.1 Introduction

11.1.1 *Hyperspectral Correlation Imaging*

Correlation imaging is a new imaging technique that is different from traditional imaging systems. It originated from the concern of the quantum fluctuations of the light field in the 1980s. In the 1990s, researchers at the University of Maryland used quantum entangled light sources for the first time to achieve an imaging mode that breaks through the limits of traditional optical diffraction and therefore is also called quantum imaging or ghost imaging [1]. The distinguishing feature of the correlation imaging is that the detection of objects and the reconstruction of images are performed separately. Light emitted from a light source is split into two correlated lights by a beam splitter: signal light and reference light. After the signal light passes through the object to be imaged, the total intensity of light passing through the various parts of the object is collected by the bucket detector, but the spatial information is not recorded. The reference light passes through without the object and the spatial distribution information of the light energy is recorded by the detector plane. Finally, the image of the target is obtained by using the correlation reconstruction algorithm of the signal light and the reference light.

As the knowledge deepens further, researchers have broken the limitations of entangled light sources and proposed using more easily realized heat sources for correlation imaging [2]. Recent studies have shown that sparse constraint theory can be used to reconstruct the target image with far less measurement data, and greatly improve the detection efficiency of the correlation imaging. Therefore, it can also be called as computational imaging or compressive imaging [3]. Due to the radical difference of the imaging target detection method and the image acquisition method, the correlation imaging has the advantage that traditional optical imaging systems cannot reach [4], mainly reflected in: (1) It can break the diffraction limit of the traditional imaging system and achieve super resolution imaging [5], (2) It can work in extremely unfavorable environments such as low light, strong fog and turbulence [6], (3) It can greatly reduce the system complexity and cost [7], (4) With extremely high security and confidentiality, it is almost impossible to correctly recover images from the stolen measurement data in the unknown measurement mode [8]. Based on these significant advantages, the correlation imaging theory has received extensive attention from researchers in recent years. The US Department of

Defense has also clearly stated that it will develop correlation imaging systems for use in future battlefield investigations.

On the other hand, hyperspectral imaging technology is a new type of remote sensing technology developed since the 1980s. It is a comprehensive technology integrating detection technology, computer technology, precision optical machinery, weak signal detection, and information processing technology. It can obtain the corresponding rich spectral information while obtaining the spatial information of the ground target, and can achieve the synchronous acquisition of spatial information, radiation information and spectral information of the ground object [9]. Compared with ordinary remote sensing methods, hyperspectral remote sensing data has characteristics of multiple bands, high spectral resolution, large data redundancy, and high spatial resolution. Hyperspectral remote sensing can make use of the diagnostic spectral features of ground objects to quantify the physical and chemical properties of the earth's surface and distinguish the features of ground objects that are often indistinguishable by multi-spectral data and plays an increasingly important role in marine remote sensing, vegetation research, geological surveys, atmospheric remote sensing and other fields. However, with the rapid development of hyperspectral remote sensing and the increasing application requirements, the problems and challenges faced are becoming more and more prominent. Hyperspectral remote sensing requires a greater amount of data to be collected and transmitted which brings difficulties to storage and transmission. Secondly, with the constant increase of resolution requirements, the scale and complexity of the corresponding imaging system will also be greatly improved. For example, the traditional optical imaging system requires a large optical system with tens of meters in diameter to achieve one-meter resolution in the geostationary orbit. The imaging system based on correlation imaging theory can precisely solve the bottleneck problem in the development of hyperspectral remote sensing because of its unique advantages. It is a potential development direction in the field of hyperspectral remote sensing.

11.1.2 Correlation Reconstruction via Sparse Dictionary Constraint

The correlation imaging system must solve the target image from the randomly-encoded measurement data through an appropriate reconstruction algorithm. Therefore, the performance of the reconstruction algorithm plays a crucial role in the imaging system. The earliest algorithms used for correlation reconstruction are first-order or higher-order correlation algorithms based on statistical theory. These algorithms calculate the values of each image point one by one. The calculation procedure is simple, and it can be updated synchronously when the measurement data is acquired. However, this type of algorithm inevitably has correlation noise. At the same time, a large amount of measurement data needs to be collected, and the image signal-to-noise ratio is difficult to increase [10]. Therefore, in recent years in the

thermal correlation imaging, researchers have increasingly used nonlinear optimization methods combined with sparse constraints to perform simultaneous calculations on all measurement data and coding modes. In particular, with the introduction and widespread application of compressed sensing theory, it provides a reliable theoretical basis for correlation reconstruction based on sparse constraints [11]. As early as the 1970s and 1980s, sparse reconstruction ideas were used in fields such as exploration and astronomical observations. In 2006, Candes et al. first mathematically deduced the conditions for the sparse signal to be completely reconstructed, and proposed the Restricted Isometry Property theory [12]. The theory shows that the number of samples required for complete signal reconstruction is proportional to the number of non-zero values of the signal and the logarithm of the total signal length. Sparse-constraint-based reconstruction method not only can effectively improve the SNR of the reconstructed image, but also can significantly reduce the amount of measurement data needed for reconstruction, which greatly improves the measurement efficiency of the correlation imaging system. Therefore, it has gradually become the mainstream method in the related imaging field.

The classical sparse constrained method in the reconstruction of the correlation image is the constraint of the discrete gradient of the target image and is also called the Total Variation constraint [13]. According to different definitions of discrete gradients, different total variation calculation models can be used. The solution to the reconstructing problem under total variational constraints can be transformed into a quadratic optimization problem.

Although the method of correlation reconstruction under total variational constraints has achieved quite good results, the method still has obvious disadvantages, the total variational constraint model assumes that all images have a very low discrete gradient and have a good uniformity. However, not all the natural images have good block uniformity, especially high-resolution remote sensing images contain very rich detailed information. These details are easily overlooked in the reconstruction of the total variation constraint model, resulting in reduced resolution of the image. In addition, the total variation model is not suitable for objects other than spatial images, such as ground spectrum.

For the reconstruction of hyperspectral images, in addition to the total variation constraints on spatial information, researchers also proposed some constraints on spectral information. There is a strong correlation between spectral segments of hyperspectral images. Thus, spectral correlation is used to propose different constraint models for hyperspectral reconstruction. The two-dimensional TV model was extended to three dimensions for hyperspectral reconstruction [14]. The property of the current reconstructed reference band was used to predict other bands [15]. And the low-rank property with the nuclear norm constraint of the three-dimensional hyperspectral data cube was applied in hyperspectral reconstruction [16]. The reconstruction model based on structure relation and researching the spectral statistical correlation is put forward [17]. Although the above model based on spectral correlation improves the reconstruction accuracy to some extent. However, the constraint model is easy to produce the problem of excessive correlation between spectra,

resulting in spectral accuracy deviation. It is more obvious for hyperspectral scenes with a wide range of spectral bands.

In the latest research of correlation reconstruction, researchers try to introduce the sparse representation of images and signals into the hyperspectral reconstruction and seek a more efficient and accurate method for complex scenes imaging. Since traditional analytically designed dictionaries, such as wavelets, curvelets, and DCT, are all on the fixed domain, that is, they are a set of fixed bases and lack the adaptability. Under normal circumstances, high sparsity cannot be achieved, and its application is limited. Therefore, how to design a self-adaptive dictionary based on the characteristics of the signals to obtain higher sparsity is crucial. In recent years, one of the main reasons for sparse representation as a research hotspot is the emergence of various adaptive dictionary learning methods, making it possible to obtain an adaptive dictionary [18].

At present, the sparse coding theory of images is mainly used in image denoising, target classification and identification, and has achieved a series of results. Duarte et al. [19] proposed a sparse coding-based classification method for the three target simulation models. Simulation results show that the recognition rate can reach 100% when the number of samples is a specific value. Wright et al. [20] applied the sparse representation theory to face recognition. Face recognition based on sparse representation achieved a good recognition effect, and the recognition accuracy was better than the traditional nearest neighbor algorithm, linear SVM algorithm, etc. Kang et al. [21] studied the optical image recognition based on the sparse representation theory and proposed a SIFT feature sparse representation based on transform domain. The extracted features can satisfy redundancy, security and compactness at the same time. Estabridis [22] uses the sparse representation for target recognition of visible and infrared images acquired by NVESD. The proposed automatic target recognition system achieves high recognition performance while reducing system complexity and improving data processing speed.

Although the image sparse representation theory has achieved remarkable results in the fields of image denoising, target recognition, etc., there is a limitation that the image to be processed has more prior knowledge in current applications, such as face recognition, vehicle identification, etc. Therefore, it is possible to construct dictionaries with strong sparsity. However, in correlation imaging, the target scene is unknown before it is reconstructed, and it is often difficult to construct a targeted sparse dictionary. Because of this, currently in the correlation reconstruction based on sparsity constraints, the fixed dictionary forms such as DCT and wavelet dictionary are mainly used. Therefore, this paper proposes a study of sparse dictionary construction methods in hyperspectral imaging, including the construction, optimization, and application of hyperspectral sparse dictionaries in correlation imaging. A set of effective hyperspectral correlation reconstruction methods based on sparse constraints is formed to make up for the lack of theoretical and applied research on hyperspectral sparseness characteristics in current correlation imaging.

The main structure of this paper is organized as follows: Sect. 11.1 is the introduction of the hyperspectral correlation imaging and the sparse constraint method. Section 11.2 presents the main methods and algorithms of sparse representation and

dictionary construction. Section 11.3 proposes the hyperspectral sparse dictionary construction method for correlation imaging. In Sect. 11.4 the dictionary selection and optimization method in the hyperspectral correlation is proposed. In Sect. 11.5 we will discuss the application of the hyperspectral sparse dictionary in real imaging system and verify the proposed method in two types of systems. And Sect. 11.6 concludes the paper.

11.2 Sparse Representation and Dictionary Construction Theory

11.2.1 Sparse Representation Theory

Obviously, the processing of sparse signals is far simpler and more efficient than non-sparse signals. This is also the reason of the widespread application of signal sparsity theory. However, in the practical application of remote sensing imaging, there are hardly natural ideal sparse signals in both the spatial scene and the spectral domain. Therefore, it is necessary to find a transformation so that the transformed signal is a sparse signal that can be processed. After the transformed sparse signal is processed, inverse transformation is performed and the original signal is returned to the domain where the final signal is processed. A linear transformation is usually used to perform sparse processing on the signal, and a linear transformation is sought for signal $x \in \mathbb{R}^N$ such that the sparsity of the transformed signal $A : \mathbb{R}^N \rightarrow \mathbb{R}^M$ satisfies the requirements. Formulated as

$$A : \mathbb{R}^N \rightarrow \mathbb{R}^M \quad s.t. \quad S(Ax) > s_0 \quad (11.1)$$

In the sparse representation of signals, the L_0 -norm is the most basic indicator of signal sparsity. In the L_0 -norm representation, the sparse representation of the signal can be described as: Finding as few vector combinations as possible from a set of base vectors so that the target signal can be used in the exact representation of the vector combination. Assume x is a signal of length N , and $D = [d_1, d_2, \dots, d_K]$ is the base vector in N dimension space. Each base vector is called an atom. The combination D of all atoms is called a dictionary. When $K = N$ the dictionary D is called a complete dictionary, $K > N$ is called an over complete dictionary or redundant dictionary. Using the atom in dictionary D to linearly represent x , that is,

$$x = \sum_{i=1}^K \alpha_i d_i \quad (11.2)$$

If x can be linearly represented by the vectors of number L in the total K basis vectors (usually L is much smaller than K and N), then x is said to be sparsely

represented by the dictionary D , and its sparsity is L .

$$x = \sum_{i=1}^L \alpha_i d_{\delta_i} + \varepsilon, \quad \delta_i \in 1, 2 \dots K \tag{11.3}$$

ε is the sparse representation error or the residual vector of the formula. The formula can be expressed in matrix form.

$$x = Ds + \varepsilon \tag{11.4}$$

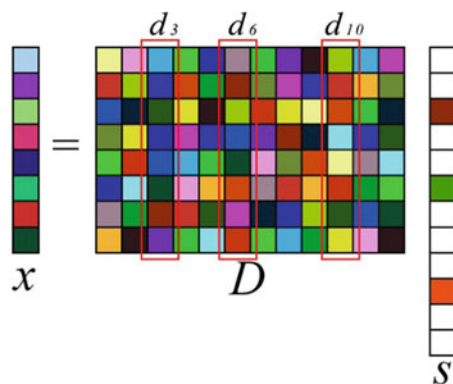
where s is a sparse vector with only L non-zero values, which called the sparse coefficient of x under the dictionary D . And $\hat{x} = Ds$ is the reconstruction signal [23]. On the premise of known signal x and dictionary D , the solution to the sparse vector s is called sparse coding or sparse decomposition. Figure 11.1 graphically illustrates the process of sparsely representing the target signal using a small number of atoms in the dictionary. The non-sparse signal x to be represented in the figure is a signal vector of length 8, the dictionary D for sparse representation contains 12 atoms, and the signal x can be represented by the linear combination of the 3rd, 6th, and 10th atoms in the dictionary D . Therefore, only the values of the 3rd, 6th, and 10th positions in the corresponding sparse vector s are nonzero, the rest of the positions are 0, and the L_0 -norm of the sparse vector s is 3.

In solving the sparse coding problem, on the one hand, we hope to sparsely represent the target signal with as few atoms as possible in the dictionary. That is, the sparsity of the sparse vector should be as high as possible. On the other hand, it is hoped that the residual vector after sparse representation is as small as possible.

Corresponding to sparsity as the direct optimization goal, the sparse coding problem can be expressed as an optimization formula:

$$\min_s S(s) \quad s.t. \quad \|x - Ds\|_2^2 \leq \varepsilon \tag{11.5}$$

Fig. 11.1 Sparse representation of the signal. x is the signal to be represented, D is the sparse dictionary and s is the sparse vector



or

$$\min_s S(s) \quad s.t. \quad x = Ds \quad (11.6)$$

Formula (11.5) and (11.6) are inequality and equality constraints, respectively.

Corresponding to the direct optimization goal with the residual, the sparse coding problem can be expressed as the optimization formula:

$$\min_s \|x - Ds\|_2^2 \quad s.t. \quad S(s) \leq L \quad (11.7)$$

In addition, sparsity and residual can be optimized at the same time, and the weight parameter λ can be used to control the optimal ratio between sparsity and residual. It can be expressed as an unconstrained optimization:

$$\min_s \frac{1}{2} \|x - Ds\|_2^2 + \lambda S(s) \quad (11.8)$$

where $\frac{1}{2} \|x - Ds\|_2^2$ is a confidence term that represents the error between the reconstructed signal and the target signal. The coefficient 1/2 is introduced to make the coefficient of the gradient after the optimization of the objective function be 1. $S(x)$ is a sparse constraint term, the sparsity function is usually L0 norm or L1 norm. As the solution of the unconstrained optimization problem is simpler, the weight parameter can be used to flexibly control the optimization ratio between the confidence item and the sparse item. The optimization model of (11.8) is more widely used in sparse coding.

In the signal sparse representation theory, two basic problems are:

- (1) How to solve the corresponding sparse vector under a given dictionary, that is the sparse coding algorithm;
- (2) How to build a dictionary to ensure its sparse representation ability of the target.

11.2.2 Sparse Coding Algorithms

In the sparse coding problem, the most direct parameter that characterizes the sparsity of a sparse signal is the L0 norm of a vector, which represents the number of atoms used for sparse representation in a sparse dictionary. The L0 norm optimization problem is usually solved by using the model of (11.6). This problem can be intuitively described as: how to select L atoms from the K atoms of the dictionary, so that the target signal is sparsely represented by the L atoms with the smallest residual.

However, since the L₀-norm is discontinuous, traditional optimization methods are difficult to optimize. The method of violent enumeration must be used, that is,

to traverse all C_K^L dictionary atom selection methods. Although this method can theoretically obtain a global optimal solution, the number of searches required by the algorithm increases exponentially as the size of the dictionary increases.

Since it is very difficult to obtain the global optimal solution, it can only be solved by using an unstable greedy algorithm. The greedy algorithm starts from the initial empty set, selects one or more atoms from the dictionary that most closely match the current sparse representation residual to join the set, and then updates the residuals to continue searching for new atoms until the number of selected atoms meets the requirements. Based on the above ideas, researchers have proposed greedy algorithms such as matching pursuit (MP) [24], orthogonal matching pursuit (OMP) [25], regularized positive matching pursuit (ROMP) [26], and segment orthogonal matching pursuit (StOMP) [27].

Although the L0 norm optimization is the most intuitive model in the sparse coding problem, the global optimal solution cannot be obtained due to the discontinuity of the L0 norm. Greedy algorithm is a way to solve L0 norm optimization but theoretically cannot guarantee correctness of solution. It is an unstable algorithm. Therefore, the researchers propose to relax the L0 norm to the L1 norm and optimize the L1 norm as a sparsity evaluation function. The sparse coding problem based on the L1 norm can be expressed as:

$$\min_s \frac{1}{2} \|x - Ds\|_2^2 + \lambda \|s\|_1 \quad (11.9)$$

where the L1 norm is the sum of the absolute values of the components in the vector.

The L1 norm is a continuous convex function, but it is not derivative. Therefore, it is not possible to directly use the classical gradient descent algorithm for optimal solution. It can be converted to a derivative function by variable substitution or approximation. Based on the gradient descent, researchers improved the calculation of the step size and search direction, such as GPSR-Basic, GPSR-BB [28], allowing the algorithm to converge at a faster rate during gradient descent.

Is it possible to find a continuous derivative smooth function approximating the L0 norm directly to solve the L0 norm optimization problem? The answer is yes. Mohimani [29] first proposed the use of Gaussian functions to approximate the L0 norm. In the experiment, it was verified that the algorithm can obtain the accuracy equal to the L1 norm optimization in a very short calculation time. Therefore, it has been increasingly used in sparse coding applications that require solving sparse constraints in recent years.

For the Lp norm optimization problem in sparse coding:

$$\min_s \|s\|_p \quad s.t. \quad x = Ds \quad (11.10)$$

Naturally to think that when $p = 2$ directly analytical solution can be given:

$$s^* = (D^T D)^{-1} D^T x \quad (11.11)$$

Equation (11.11) is the so-called least-squares solution, and the L2 norm optimization for Lagrange multipliers:

$$\min_s \|x - Ds\|_2^2 + \lambda \|s\|_2^2 \tag{11.12}$$

There exists analytic solution:

$$s^* = (D^T D + \lambda I)^{-1} D^T x \tag{11.13}$$

Equation (11.12) is also called the LASSO problem. Although there are analytical solutions for LASSO problem, it is difficult to guarantee the sparsity of solutions in sparse coding problems. To solve the above problem, the FOCUSS algorithm is proposed by adding weights on the sparse signal in each iteration [30].

Define the weight matrix

$$W = \text{diag}(w_i) \tag{11.14}$$

Calculate the transformation of the signal under the weight matrix,

$$s = Wq \tag{11.15}$$

Equation (11.12) can be expressed as

$$\min_q \|x - DWq\|_2^2 + \lambda \|q\|_2^2 \tag{11.16}$$

The solution is

$$s^* = Wq^* = WW^T D^T (DWW^T D^T + \lambda I)^{-1} x \tag{11.17}$$

Use the last result to update the matrix of weights and continue to iterate into the above formula until it meets the convergence conditions.

In addition, it is also possible to approximate the Lp norm optimization problem by adding the weight $w_i = |s_i^{(n-1)}|^{p-2}$ to each unknown in the least-squares problem, where n is the number of iterations. This strategy can continuously approximate the Lp norm solution in iterations. This method is called the iteratively reweighted least squares method (IRLS) [31].

Five sparse coding methods are used in this paper for comparison: OMP, SLO, GPSR, FOCUSS and IRLS.

11.2.3 Sparse Dictionary Construction Theory

There are two core issues in the sparse representation of signals, one is sparse coding to solve the sparse representation of signals under a given dictionary, and the other is the construction of sparse dictionaries. For the sparse coding problem, there have been more than 20 years of research history, the theory and method are relatively complete, and the research on the more basic sparse dictionary related theory is relatively lacking. In recent years, it has gradually formed a theoretical system. The types of dictionary mainly contain the fixed dictionary, the sample dictionary and the training dictionary.

Fixed dictionary refers to the use of certain mathematical rules to generate atoms in the dictionary and has no correlation with the target signal. It has universal applicability. The normal fixed dictionaries are DCT dictionary, wavelet dictionary, differential dictionary or Gabor dictionary.

The atoms of the sample dictionary are the sample signal that is similar to the target signal type, the signal can be constructed by actual measurement data or simulation data. This is a kind of dictionary with a priori knowledge, which requires a certain understanding of the target signal in advance. The sample dictionary is generally over-complete and has a large degree of redundancy. The size of the high-dimensional signal dictionary may be quite large. Therefore, in some areas such as target recognition, the samples are often processed by means of dimensionality reduction, and samples such as PCA, SIFT, etc. are extracted as atoms of the dictionary. The formed dictionary is called a feature dictionary, and then the target signal is also extracted from the corresponding feature and is sparsely coded under the feature dictionary, and the identification and matching are performed according to the encoding result. Sparse coding-based feature dictionary is an effective recognition method. However, in some other fields, such as the correlation imaging studied in this paper, the feature dictionary is not applicable because the imaging system directly encodes the scene itself.

The training dictionary is used to perform machine learning on the sample set (i.e. sample dictionary or feature dictionary in the previous section) to generate a new dictionary so that all sample signals in the sample set can be sparsely represented by the new dictionary. Usually the number of atoms contained in the new dictionary is much smaller than the number of training samples. The process of dictionary learning is to shrink the sample set. On the one hand, the training dictionary can construct targeted sparse dictionaries based on the prior features of the signal, and on the other hand, it avoids the huge computational burden of directly adopting the sample dictionary. Therefore, it has gradually been applied by more and more researchers and is also the focus of this article.

The dictionary training process is described by the following mathematical model. The input sample set matrix $W = [w_i]_{i=1}^L$ is arranged from the column vectors w_i of the L training sample signals. The targeted dictionary to train is $D = [d_k]_{k=1}^K$, K is the number of atoms in the dictionary to be trained. The sparse coding coefficient of the sample set under dictionary D is $X = [x_i]_{i=0}^L$. The dictionary training problem

is to solve the following optimization problems.

$$\min_{D, X} \sum_{i=1}^L S(x_i) \quad s.t. \quad W = DX \quad (11.18)$$

The condition number is NL (N is the dimension of the space where the signal is located), and the unknown quantity to be solved is $(N + L)K$. Examine the morbidity of the problem, set $K = \alpha N$ and $L = \beta N$, usually $L \geq K > N$. The morbidity is defined as the ratio of the unknown to the condition number.

$$\eta = \frac{K(N + L)}{NL} = \alpha \left(1 + \frac{1}{\beta}\right) \approx \alpha \quad (11.19)$$

The training problem for the redundant dictionary ($\alpha > 1$) is ill-conditioned, and the higher the redundancy of the dictionary, the more ill-conditioned the problem is. Therefore, it is difficult to find the global optimal solution. Usually, the greedy iterative idea is used to find the local optimal solution. The main strategy is: (1) Update sparse coding coefficients under a fixed dictionary, that is the sparse coding problem discussed before. (2) Update dictionary using sparse coding coefficients. According to different dictionary update algorithms, researchers proposed dictionary learning algorithms such as MOD, ILS-DLA, ODL, and K-SVD. These four training methods will be used in the future sections.

K-SVD (K-mean Singular Value Decomposition) algorithm is currently recognized as one of the best comprehensive dictionary learning algorithms and has been widely applied in sparse dictionary construction [32]. The K-SVD algorithm updates the atoms in the dictionary one by one. For the i -th atom, first calculate the residual matrix after removing the i -th atom from the dictionary according to the sparse coding coefficients.

$$E = W - \sum_{j \neq i} d_j x_j \quad (11.20)$$

Then the singular value decomposition of the residual matrix is calculated,

$$E = U \Sigma V^T \quad (11.21)$$

Update the i -th atom in the dictionary as the first column vector of the matrix U , and update the i -th sparse coding coefficient as the product of the largest singular value and the first column vector of the matrix V .

11.3 Hyperspectral Sparse Dictionary Construction Method for Correlation Imaging

11.3.1 Sparse Representation in Hyperspectral Correlation Imaging

Different from the “shoot and gain” characteristic of traditional imaging systems, the significant feature of the correlation imaging is that the detection of objects and the reconstruction of images are performed separately. The traditional optical observation is based on the measurement of the distribution of the intensity of the light field. There is a one-to-one relationship between the detection target’s pixel and the detector’s pixel. The correlation imaging is based on the correlation measurement of the intensity of the light field. The relationship between the detection target pixel and the detector pixel is a many-to-one relationship. The correlation imaging system performs a series of correlation measurements through a series of uncorrelated coding patterns and reconstructs the target image from the measurements sequence and the corresponding measurement patterns through the reconstruction algorithm.

Assume the column vector y with length m is the m -times correlation measurements. All the measurement patterns can be represented as the matrix A of size $m \times n$. Matrix A is also called a measurement matrix, where the i -th row of the matrix represents the measurement mode of the i -th correlated measurement, and n represents the unknown quantity to be solved, that is the total number of scene pixels. x represents a one-dimensional expansion vector of the target scene, and the correlation measurement process is represented as

$$y = Ax + e \quad (11.22)$$

where e is the unavoidable error and noise item in the measurement process.

For hyperspectral remote sensing correlation imaging, set the number of spatial pixels in the imaging scene to be $n_1 \times n_2$ and the number of spectral bands to n_3 . According to different measurement modes, target x has the following three forms:

- (1) x is a spatial image of a spectral band, which is a column vector of length $n_1 n_2$. Perform the same spatial coding for each spectral band, and the number of n_3 correlation reconstructions is required.
- (2) x is the spectral curve of a spatial position, which is a column vector of length n_3 . The same spectral encoding is performed for each spatial location and the number of $n_1 n_2$ correlation reconstructions is required.
- (3) x is a one-dimensional expansion of the entire hyperspectral scene, which is a column vector of length $n_1 n_2 n_3$. Spatial and spectral domains are mixed and coded simultaneously, with only one correlation reconstruction.

The above measurement modes are defined as spatial correlation measurement, spectral correlation measurement, and joint correlation measurement, respectively. The spectral correlation measurement model studied in this paper is the focus, while

considering the other two models. In the description and derivation of the theory, all three models are applicable unless otherwise specified.

Introduce sparse representation theory if the target scene x can be sparsely represented by a small number of atoms in dictionary D , i.e.

$$x = Ds \quad (11.23)$$

Substituting into Formula (11.23), the correlation measurement can be expressed as

$$y = ADs + e \quad (11.24)$$

The problem of reconstructing the target scene x is transformed into solving the sparse coefficient s , corresponding to the optimization problem.

$$\min_s \frac{1}{2} \|y - ADs\|_2 + \lambda S(s) \quad (11.25)$$

Comparing the above formula with the Formula (11.8), we can find that the two are very similar, but in Formula (11.25), there is one more measurement matrix A before the dictionary D . Define the matrix AD as the sensing matrix and an important conclusion can be derived:

Theorem The reconstruction of correlation images under the sparse representation model is equivalent to the sparse coding problem under the sensing matrix.

Therefore, in addition to the dictionary training algorithm, the sparse encoding algorithm discussed earlier is also a reconstruction algorithm of the correlation imaging system.

The ratio of the number of measured data to the number of reconstructed image pixels is defined as the sampling rate.

$$sr = m/n \quad (11.26)$$

This article studies the under-sampling problem of the sampling rate $sr < 1$. The lower the sampling rate, the more efficient the correlation image acquisition data is, but the more difficult the reconstruction is.

The ratio of the number of measured data to the number of reconstructed sparse coefficients is defined as coding redundancy.

$$cr = m/k \quad (11.27)$$

The higher the coding redundancy, the more atoms that can be selected by the reconstruction algorithm in the dictionary, but the harder it is to guarantee the sparsity of the reconstruction.

11.3.2 The Fixed and Sample Hyperspectral Sparse Dictionary Construction

This section and the following section focus on the spectral correlation measurement system, constructs a fixed dictionary, a sample dictionary and a training dictionary, and performs spectral correlation measurement and reconstruction experiments. The experimental results are compared and analyzed.

The USGS_MIN spectral library is used as a dictionary for training samples and test samples. The USGS_MIN spectral library contains 403 spectral data, each of which contains data points for a total of 420 spectral segments in the 0.395–2.56 μm wavelength range. 300 of these lines are used as training data and the rest is used as test data.

First a fixed dictionary is built, which does not require any prior information including training samples. The DCT dictionary can be used in spectral correlation reconstruction, but the traditional orthogonal DCT dictionary for spectral correlation reconstruction contains too many high-frequency atoms. Reduce the maximum frequency of atoms in the DCT dictionary, the dictionary atom is constructed as

$$d_{i,j} = \cos\left[\frac{\pi}{a}i\left(j + \frac{1}{2}\right)\right] \quad (11.28)$$

The down frequency DCT dictionary no longer has orthogonality. Set the down-shift ratio a/N to 0.05. As shown in Fig. 11.2, the 20th to 30th atoms in the DCT dictionary before and after the down frequency are ranked from low to high, and the down frequency atoms are much smoother.

Experiment 1 Spectral correlation measurement and reconstruction of Zoisite spectral samples in the USGS_MIN spectral library is performed. In the experiment, the measurement matrix is an equal probability $(-1,1)$ Bernoulli random matrix with a sampling rate of 0.1. Add random noise to the measurement signal results in a signal-

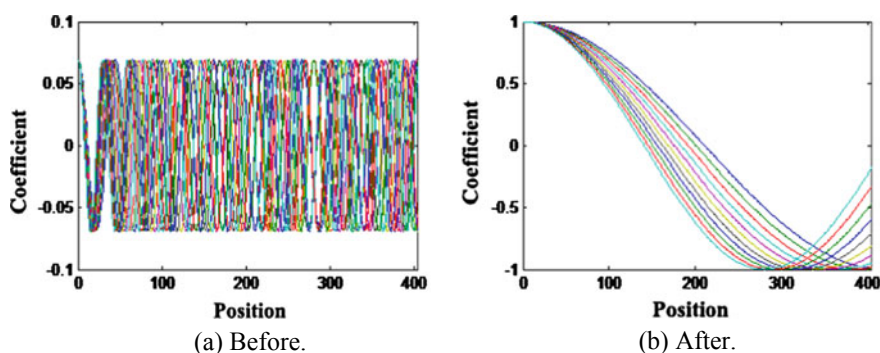


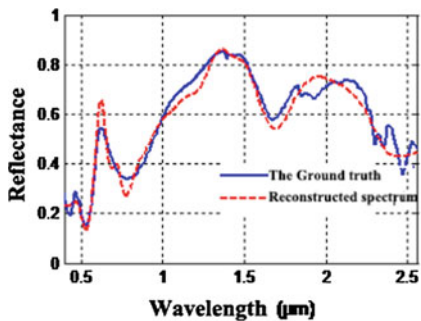
Fig. 11.2 The 20–30th atoms are arranged by frequency in the DCT dictionary before and after the down frequency

to-noise ratio of 40 dB (The noise signal energy is about 1% of the measurement signal). Five sparse coding algorithms are used for spectral reconstruction under the fixed down frequency DCT dictionary, and their RMSE errors are calculated. The algorithm parameters are set as followed. The number of iterations in the OMP algorithm is 20. The Lagrange multiplier in the GPSR algorithm is 10. The norm p in the IRLS algorithm is 0.5 and the maximum number of iterations is 50. The SL0 algorithm has a descent step of 2, a shrinkage factor of 0.5, and a parameter σ convergence of 0.04. In the FOCUSS algorithm, the power p is 0.6 and the Lagrange multiplier is 0.01. The setting of the above parameters is a parameter that is better in the appropriate calculation time and summed up in the empirical values under multiple experiments. The reconstruction results are shown in Fig. 11.3.

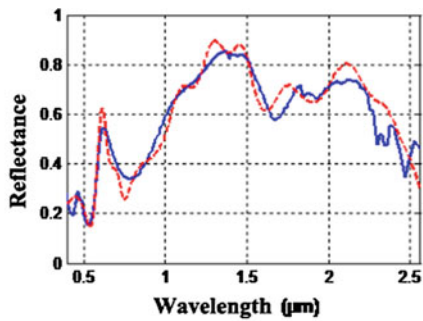
Spectral curves of 300 samples in the USGS_MIN spectral library are constructed as a sample dictionary. Each atom in the dictionary is a spectral curve in the spectral library. Some of the spectral samples in the dictionary are shown in Fig. 11.4. It can be seen that the dictionary contains various spectral samples with significant differences.

Experiment 2 The spectral sample dictionary is used in spectral correlation reconstruction. The experimental environment and parameters were the same as in Experiment 1. The reconstruction of the spectrum under different reconstruction algorithms is shown in Fig. 11.5. Compared with the improved results of the down frequency DCT dictionary reconstruction (Fig. 11.3), the spectral reconstruction error of the sample dictionary is larger than that of the down frequency DCT dictionary under all algorithms. The most significant difference is the GPSR algorithm (RMSE of the down frequency DCT dictionary = 5.05% and the RMSE of the spectral sample dictionary = 13.3%). The smallest difference is the SL0 algorithm (RMSE = 6.13% for the down frequency DCT dictionary and RMSE = 6.24% for the spectral sample dictionary). It can be seen that although in the same spectral library, due to the large differences between the different spectral curves, the atoms in the dictionary that correspond to the target features are few and the irrelevant interference atoms are too much. Therefore, the direct use of the sample in the spectral library as a dictionary does not yield ideal results in spectral correlation reconstruction.

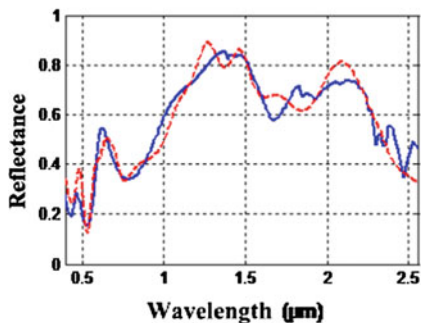
Experiment 3 Different numbers of spectral samples are used as the sparse dictionary for spectral reconstruction and the reconstruction errors are compared in this experiment. The spectral sparse dictionary is constructed by selecting different number of spectral samples from the sample library under the same conditions as in Experiment 2. Taking IRLS reconstruction algorithm as an example, spectral correlation measurement and reconstruction experiments are performed at sampling rates of 0.05, 0.1, and 0.15, respectively. We perform 100 independent randomized experiments and take the mean RMSE of each experiment. The error of spectral reconstruction in different dictionary atom numbers is shown in Fig. 11.6. It can be seen that when the sampling rate is 0.05, the reconstruction error decreases as the number of atoms included in the dictionary increases, and the reconstruction error does not substantially change with the number of dictionary atoms when the



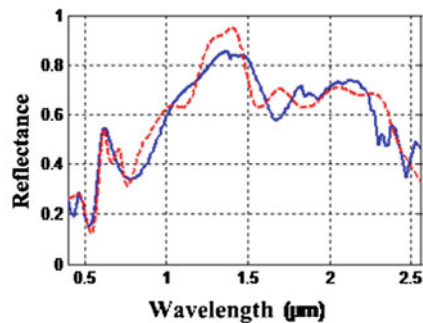
(a) OMP algorithm (RMSE=4.09%)



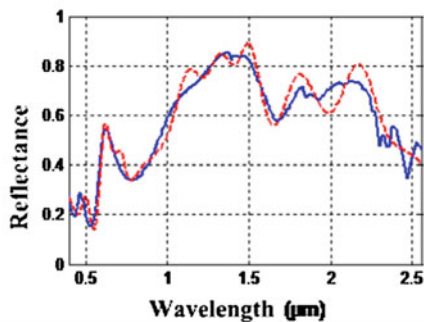
(b) GPSR algorithm (RMSE=5.05%)



(c) IRLS algorithm (RMSE=5.63%).



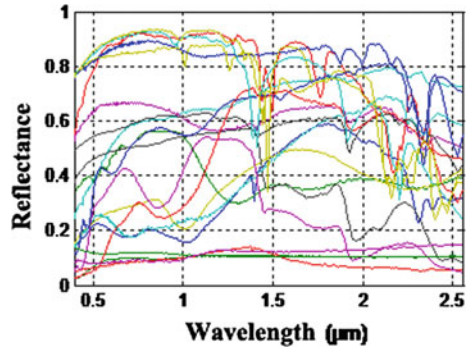
(d) SL0 algorithm (RMSE=6.13 %).



(e) FOUSS algorithm (RMSE=4.62 %).

Fig. 11.3 The spectral curve of Zoisite and the reconstruction results using the down frequency DCT dictionary under different reconstruction algorithms

Fig. 11.4 Part of the atoms in the sample spectral dictionary



sampling rate is 0.1 and 0.15. However, when the sampling rate is 0.05, the reconstructed spectral error is very large. In this case, the conclusion obtained is of little practical significance. It can be considered that the number of atoms contained in the USGS_MIN spectral library for the spectral sample dictionary is independent of reconstruction accuracy.

11.3.3 Training Dictionary Construction

The training dictionary uses a number of samples far more than the number of dictionary atoms to obtain a sparse dictionary capable of sparsely representing these samples through a machine learning algorithm. For the image processing field currently applied sparse representation theory, it is easy to obtain many spatial block samples for dictionary training. For the spectral library, the number of samples is very limited, the first problem is how to construct a large number of samples with different characteristics from a limited spectral library for training. Thus, in this paper several mathematical methods to produce more training samples of the spectral library are applied.

Defining the following operation for a spectral sample $w^{(0)}$ of length n to produce a new sample w' .

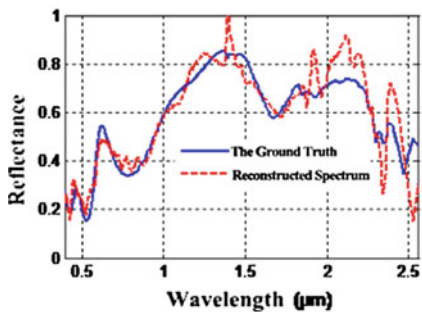
- (1) Stretch and shrink:

$$w^{(1)} = T_{\alpha}(w^{(0)}) \tag{11.29}$$

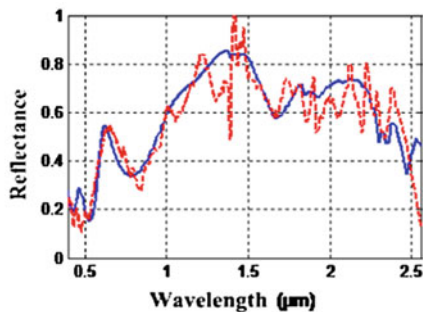
Defining the stretch factor $\alpha > 0$, The stretching operation stretches or shrinks a sample $w^{(0)}$ of length n into a sample $w^{(1)}$ of length αn by an interpolation algorithm.

- (2) Shift:

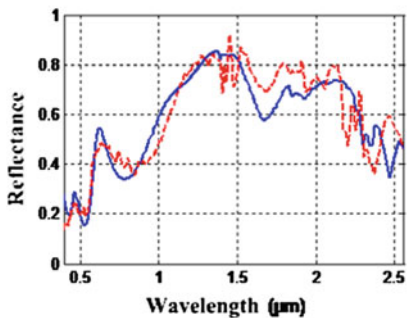
$$w^{(2)} = H_d(w^{(1)}) \tag{11.30}$$



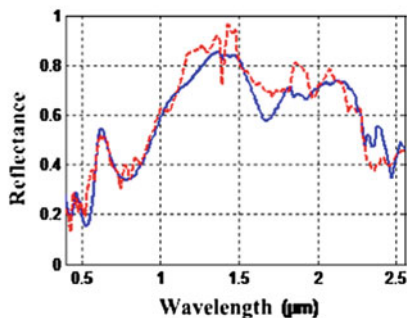
(a) OMP algorithm (RMSE=6.13%).



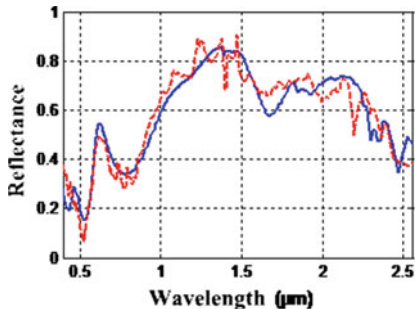
(b) GPSR algorithm (RMSE=13.3%).



(c) IRLS algorithm (RMSE=6.83%).



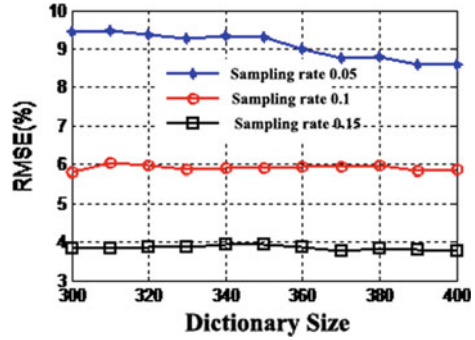
(d) SL0 algorithm (RMSE=6.24%).



(e) FOCUSS algorithm (RMSE=5.78 %).

Fig. 11.5 The spectral curve of Zoisite and the reconstruction results using the sample dictionary under different reconstruction algorithms

Fig. 11.6 The RMSE errors of the reconstructed spectrum in different dictionary atom numbers of the spectral sample dictionary



The shift amount $d \in (-\alpha n, \alpha n)$, The translation operation shifts the sample $w^{(1)}$ by d spectral bands, and the portion outside the coordinate range is discarded.

(3) Truncate and fill:

$$w^{(3)} = C_t(w^{(2)}) \tag{11.31}$$

For the stretching factor $\alpha > 1$, the length of the sample after stretching is greater than n . Intercept n pieces of spectral data from position $t \in [0, (\alpha - 1)n]$.

For the stretching factor $\alpha < 1$, the length of the sample after stretching is less than n . After the shift, the remaining spectral bands are filled, and the filled value is equal to the spectral sample edge value.

After the three steps of transformation, we have

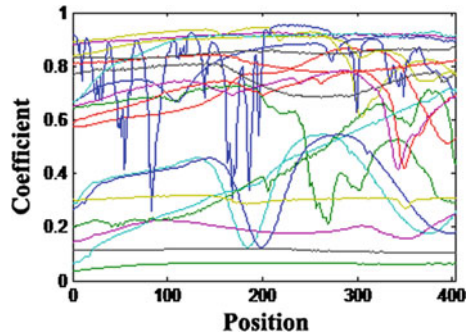
$$w' = C_t H_d T_\alpha(w^{(0)}) \tag{11.32}$$

Randomly select the training spectral sample in the sample library and set the random number between (0, 2) to be the stretching coefficient, stretching, shifting and truncating the selected samples in sequence to form a new training sample. Construct a spectral sample training set with a sample size of 10,000. Some samples in the training set is shown in Fig. 11.7. It can be seen that the training set contains both smooth samples and fluctuating samples.

Spectral sparse dictionary from spectral training samples is constructed using four dictionary training algorithms (MOD, LS-DLA, ODL, KSVD). All algorithms use high-efficiency OMP algorithms for sparse coding. The number of iterations of the OMP algorithm is 10, the number of dictionary updates is 50, and the number of dictionary atoms to train is 600. Set the forgetting factor to 0.5 for LS-DLA algorithm. The initial dictionary is set to take the first 600 samples of the training sample set.

Observe the convergence of errors with dictionary update times in dictionary training under different algorithms. The Frobenius norm is defined to be the dictionary error as the difference between the sample matrix and the product of the dictionary and the sparse coefficient matrix.

Fig. 11.7 Some of the training samples in the expanded spectral training set



$$e = \|W - DX\|_F \tag{11.33}$$

The dictionary error under different dictionary training algorithms changes with the number of updates as shown in Fig. 11.8. It can be seen that except that the ODL algorithm converges very slowly, the convergence speed of the other three algorithms is very close, while the training error of KSVD algorithm is slightly lower than that of LS-DLA and MOD algorithm. Consider the calculation time of different algorithms, the MOD algorithm takes 859 s, the ODL algorithm takes 756 s, the LS-DLA algorithm takes 1090 s, and the KSVD algorithm takes 2531 s. sothoThe KSVD algorithm takes the longest time and is 2–3 times than that of other algorithms because of the SVD decomposition calculation of the matrix for KSVD algorithm.

Experiment 4 Compare the best KSVD dictionary in the training dictionary with the down frequency DCT dictionary and sample dictionary constructed in Sect. 11.3.2 under different sampling rate and noise conditions. The remaining experimental parameters are the same as the previous experiment. Figure 11.9 shows the average RMSE error reconstructed in 100 times for three spectral sparse dictionaries when the sample rate is 0.04–0.2 with 40 dB noise. It can be seen that the spectral average error reconstructed by the KSVD training dictionary is lower than the spectral error reconstructed by the down frequency DCT dictionary and the sample dictionary,

Fig. 11.8 Variations of training error of different dictionary training algorithms with iterative times

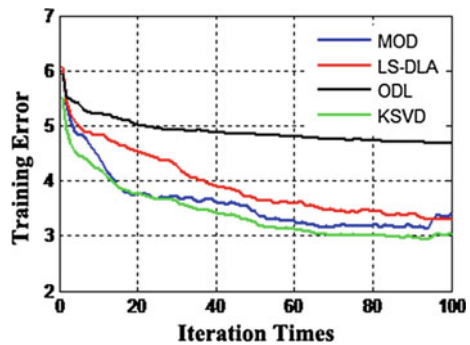
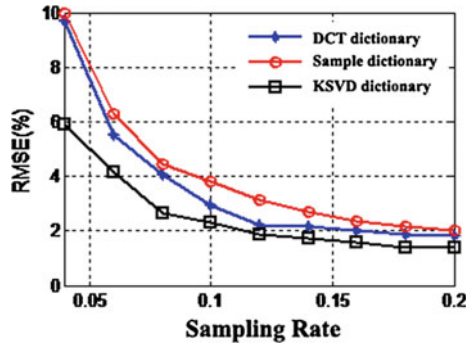


Fig. 11.9 Reconstruction errors of three spectral sparse dictionaries with different sampling rates



and the error is reduced more significantly when the sampling rate is low (less than 0.1). Figure 11.10 shows the mean RMSE error of the three spectral sparse dictionaries in 100 experimental reconstructions when the signal-to-noise ratio of the measurement signal varies from 20 to 60 dB. It can be seen that the average spectral error reconstructed by the KSVD training dictionary is still lower than the spectral error reconstructed by the down frequency DCT dictionary and the sample dictionary, and under the same SNR condition, the reconstruction error of KSVD dictionary is about 0.01–0.03 lower than that of other two dictionary reconstruction errors. As shown in Fig. 11.11, the total time for 100 times reconstruction of three spectral sparse dictionaries when the sampling rate is changed from 0.05 to 0.2 is presented. The reconstruction time required for training a dictionary with KSVD is the longest, which is about 10–20% more than that of the other two kinds of dictionary reconstruction time. This is because the KSVD dictionary is a redundant dictionary and contains a large number of atoms.

In summary, under different sampling rates and noise environments, the accuracy of the spectral sparse dictionary reconstruction based on the KSVD dictionary training algorithm is higher than that of the other two dictionaries, but the required computing time is slightly longer.

Fig. 11.10 Reconstruction errors of three spectral sparse dictionaries for different signal-to-noise ratios of measurement signals

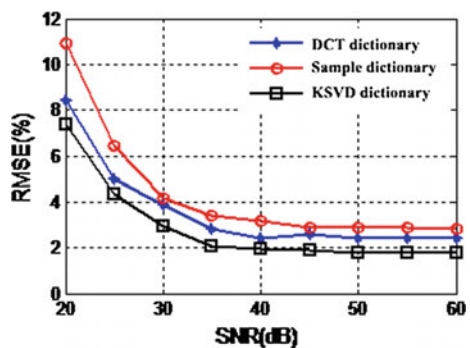
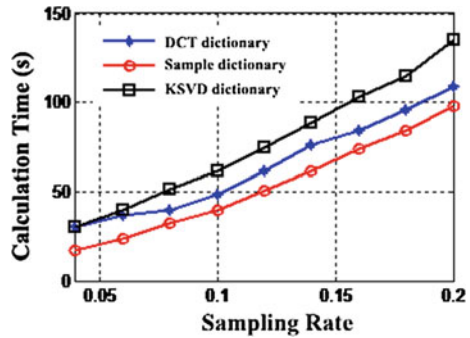


Fig. 11.11 Total reconstruction time of three spectral sparse dictionaries with different sampling rates



11.3.4 Hyperspectral Sparse Dictionary Construction of Different Class

Each ground object has its own spectral characteristics, and the unique spectral features are the theoretical basis for the traditional method of identifying the material properties of the ground features by hyperspectral remote sensing. For hyperspectral correlation reconstruction, the spectral sparse dictionary used in the reconstruction should include as much as possible the characteristics of the reconstructed target spectrum, with as few features as possible containing other types of spectrum. For these purposes, sparse dictionaries need to be trained from samples with similar characteristics.

The basis of constructing the classification spectrum sparse dictionary is the classification of spectral training samples. The most direct classification method is to classify according to the natural attributes of the target, such as iron ore, copper ore, poplar leaves, corn leaves, etc. However, this type of operation will result in a very large number of classifications, and it is impossible to include such comprehensive samples in the spectral sample database. In fact, the spectral characteristics of poplar leaves and corn leaves are highly similar, and they can be divided into one category completely. However, a wide variety of ores naturally have similar characteristics and can be grouped into one category. In this paper, clustering algorithms in machine learning are used to cluster different types of spectral samples in the spectral library to build different types of sample libraries.

K-means clustering algorithm is an effective algorithm in unsupervised clustering. The basic idea is to randomly select K samples as the initial centroid, sequentially classify each sample as the centroid closest to the sample, and then recalculate the centroids of each class until all centroids do not change.

From the USGS, ASTER, JPL, JHU and other spectral libraries, 892 different types of spectral samples are selected to build the classification database. Because the spectral range and wavelength of the spectrum in different spectral sample libraries are not the same, the cubic spline interpolation is performed for all samples in the

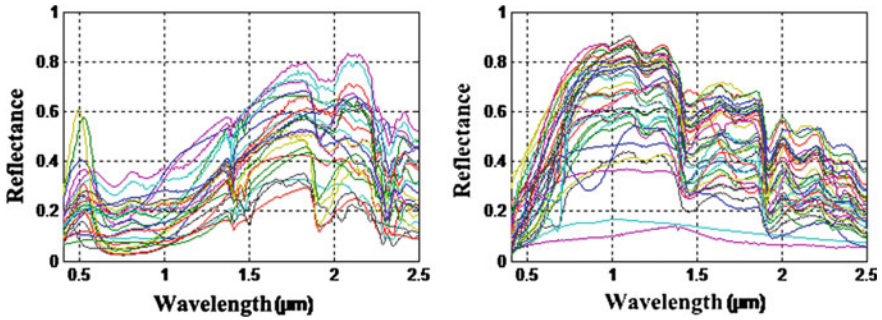


Fig. 11.12 Spectral samples from Category 1 and 2 after clustered in the sample libraries

spectral library to form spectral data of 400 spectral bands in the range of 0.4–2.5 μm wavelengths. K-means clustering is performed on the processed sample set, and the number of classifications is set to 20. As shown in Fig. 11.12, the spectral samples in the Category 1 and Category 2 sample libraries, it can be seen that the spectral samples in the same category have many similar characteristics.

Spectral sample expansion and sparse dictionary training are performed respectively on the clustered spectral sample libraries. After the expansion, each type of spectral sample library contains 10,000 samples. The KSVD training algorithm is used for dictionary learning to obtain 20 different categories of spectral sparse dictionaries. At the same time as a comparison, all types of samples in the spectral library are trained together to obtain a mixed dictionary. The sparse coding algorithm in dictionary learning is fast OMP algorithm, sparse coding iteration times is 8, and dictionary update times is set to 50.

Experiment 5 Spectral sparse dictionaries of different categories are used to sample and reconstruct spectral targets. The test samples are Jarosite in the USGS_MIN spectral library and Red Willow sample in the USGS_VEG spectral library. First, the K-nearest neighbor method is used to classify the spectrum of the test sample. The K-nearest neighbor classification idea is to find the K nearest neighbors in the training set and take the most of the K nearest neighbors as the test sample category. Here we set $K = 9$. The result of the classification is that Jarosite falls into 10th category and Red Willow falls into 11th category. Spectral correlation reconstruction using a category dictionary consistent with the test sample classification is applied. As a comparison, experiments are performed under the same conditions using several dictionaries and mixed dictionaries that does not match. The experiment uses an equal probability $(-1, 1)$ random Bernoulli measurement matrix with a sampling rate of 0.1, a measured signal-to-noise ratio of 40 dB, and uses IRLS algorithm as the reconstruction algorithm.

Figures 11.13 and 11.14 show the result of spectral correlation reconstruction of Jarosite and Red Willow under different categories of spectral dictionary. The result shows the reconstruction errors under the category 10 and 11 spectral dictionaries that are consistent with the target category are the least, and the errors under the mixed

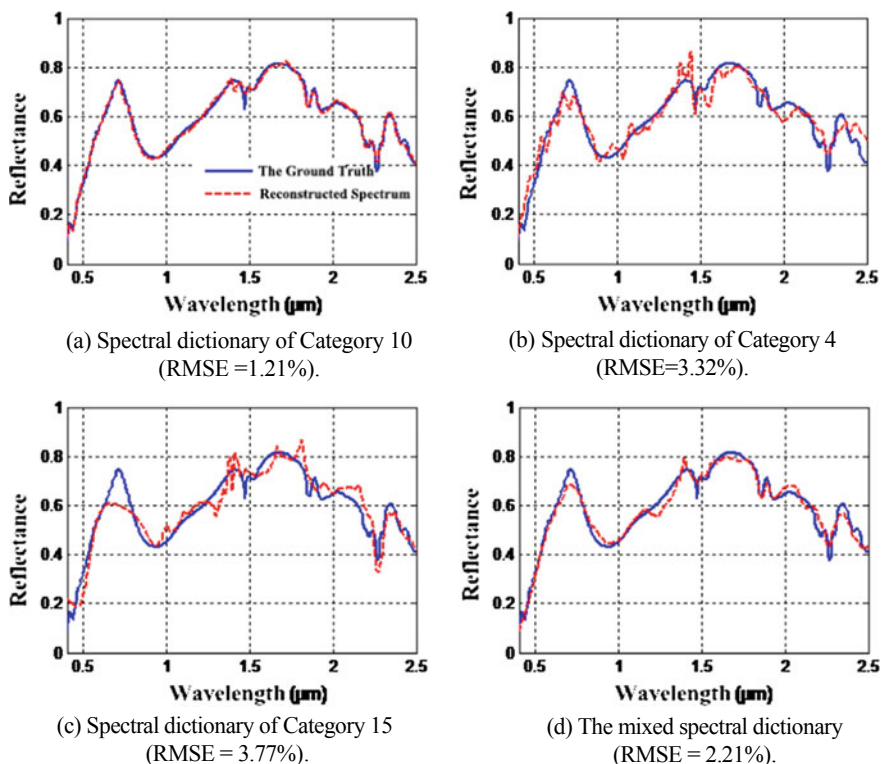
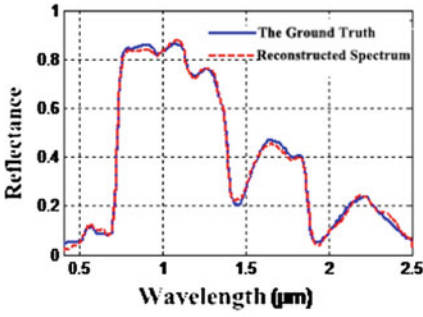


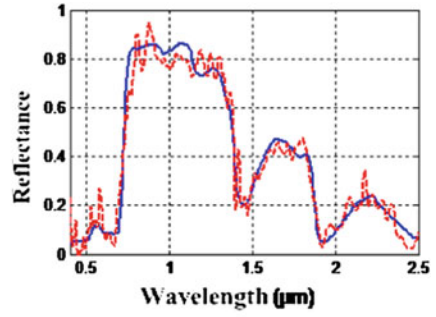
Fig. 11.13 The spectral correlation reconstruction results of Jarosite under different categories of spectral dictionary

spectral dictionary are the second, while there is a considerable amount of error in the reconstruction under the classification spectral dictionary that is inconsistent with the target type. The experiment shows that the strategy of training by the classification spectrum dictionary based on sample clustering and selecting the category dictionary matching the target can improve the spectrum sparsification ability and improve the spectral precision in the correlation reconstruction.

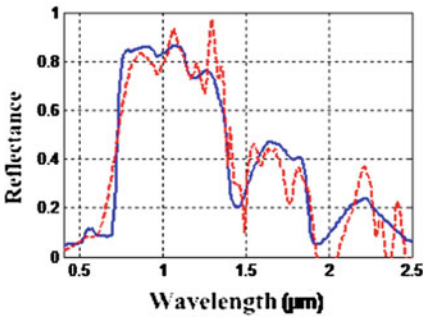
Compare the optimal classification spectrum sparse dictionary with the mixed spectrum dictionary under different sampling rate and noise conditions. The sampling rate is changed from 0.02 to 0.2 at 40 dB SNR and the signal to noise ratio is changed from 20 to 60 dB at 0.1 sample rate. Calculate the average RMSE error of two dictionaries reconstructed in 100 experiments. Figures 11.15 and 11.16 are the calculation results of the Jarosite spectrum and the Red Willow spectrum, respectively. It can be seen that for both the Jarosite spectrum and the Red Willow spectrum, under the same sampling rate and signal-to-noise ratio, the error in the spectral correlation reconstruction using the optimal classification dictionary is generally lower than that of the mixed dictionary without classification, and the advantage of using



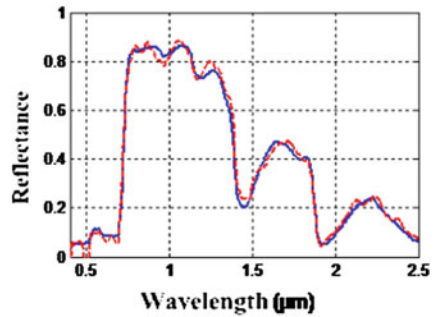
(a) Spectral dictionary of Category 11 (RMSE=1.31%).



(b) Spectral dictionary of Category 4 (RMSE=6.81%).

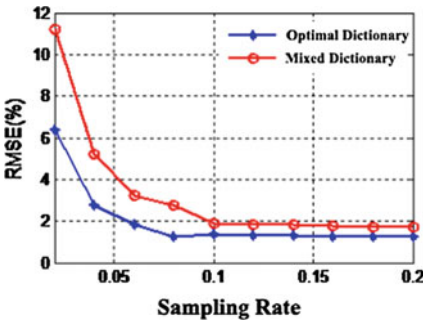


(c) Spectral dictionary of Category 15 (RMSE=8.59%).

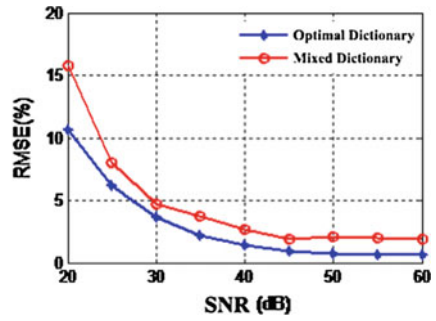


(d) The mixed spectral dictionary (RMSE = 2.21%).

Fig. 11.14 The spectral correlation reconstruction results of Red Willow under different categories of spectral dictionary



(a) RMSE errors at different sampling rate.



(b) RMSE errors at different SNR.

Fig. 11.15 The correlation reconstruction RMSE errors of Jarosite spectrum at different sample rate and signal-to-noise ratio

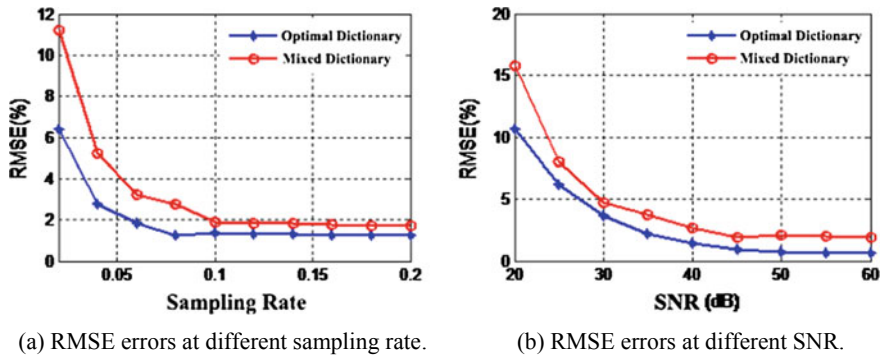


Fig. 11.16 The correlation reconstruction RMSE errors of Red Willow spectrum at different sample rate and signal-to-noise ratio

a category dictionary for reconstruction when the sampling rate is low (below 0.1) is even more significant.

11.4 Sparse Dictionary Selection and Optimization Method in the Reconstruction

The previous chapter discussed in detail the construction method of hyperspectral sparse dictionaries and achieved efficient sparsity of hyperspectral targets through the construction of spectral sample libraries and the training of classified sparse dictionaries. In the correlation reconstruction, in order to obtain a better hyperspectral target sparse effect, and to take full advantage of the spectral training dictionary, the sparse dictionary and the reconstruction method need to be adaptively improved, including different types of dictionary selection methods in the reconstruction, methods for selecting different atoms in a dictionary, and methods for dynamically updating sparse dictionary according to target features. In this chapter, based on the hyperspectral correlation sparse dictionary constructed in the last chapter, we study the improved strategies in the correlation reconstruction of the above problems, and apply the hyperspectral sparse dictionary better to the correlation reconstruction.

11.4.1 Dictionary Adaptive Selection Method in Correlation Reconstruction

Based on the clustering features of the spectrum, different categories of spectral dictionaries have been constructed and verified that dictionaries of suitable categories can achieve better reconstruction effects. However, a key issue in practical

applications is that the reconstructed target spectrum is not known in advance, and it is not possible to choose the appropriate a priori classification for the target. At the same time, there is no comparison of true values, and it is not possible to try one by one and then choose the best result.

For the selection of non-a priori classified spectral dictionaries, a preliminary reconstructed spectrum is firstly obtained by using a combination dictionary to extract partial atoms from all categories of spectral dictionaries for reconstruction. Then, by reclassifying the preliminary reconstructed spectrum and selecting the matching spectral dictionary that matches them, the measurement data is reconstructed again, and the optimized reconstructed spectral results are obtained. For the method of classification of preliminary reconstructed spectrum, this paper proposes two strategies. First, the K -nearest neighbor classification method is used to classify the preliminary reconstructed results in the known class of training samples on the premise of known spectral training samples. Second, when the spectral training sample is unknown, the preliminary reconstruction results are sparsely represented using all categories of spectral dictionaries. The GINI coefficient is used to calculate the sparseness of sparse signals in each dictionary, and the dictionary with the largest GINI value is selected. The detailed steps are as follows:

Algorithm 1: Sparse Dictionary Adaptive Selection Algorithm Based on K -Nearest Neighbor Classification (Training Samples Known) Input: The correlation measurements y , the measurement matrix A , different types of spectral sparse dictionary $D = [D_1, D_2, \dots, D_K]$ and the training spectral library $W = [W_1, W_2, \dots, W_K]$.

Step 1: Building a mixed spectral dictionary. Selecting m atoms randomly from each spectral dictionary D_i to constitute a mixed spectral dictionary D_0 .

Step 2: Preliminary reconstruction. The spectral spectrum D_0 is used to reconstruct the spectrum and the initial solution x_0 is obtained.

$$\hat{s} = \arg \min_s \frac{1}{2} \|y - AD_0s\|_2 + \lambda S(s) \quad (11.34)$$

$$x_0 = D_0\hat{s} \quad (11.35)$$

Step 3: Calculate the distance between the reconstructed spectrum x_0 and all samples in the sample set W .

$$dis[i] = \cos^{-1} \frac{\langle x_0, w_i \rangle}{\|x_0\| \|w_i\|} \quad (11.36)$$

Step 4: Category determination. The statistical distance from the nearest k samples in the sample library corresponds to the classification c , $c \in [1, 2, \dots, K]$.

Step 5: Final reconstruction. Select dictionary D_c to reconstruct the object again.

$$\hat{s} = \arg \min_s \frac{1}{2} \|y - AD_c s\|_2 + \lambda S(s) \quad (11.37)$$

$$\hat{x} = D_c \hat{s} \quad (11.38)$$

Step 6: Output the final result \hat{x} .

Algorithm 2: Sparse dictionary adaptive selection algorithm based on sparsity estimation (training samples unknown) Input: The correlation measurements y , the measurement matrix A , different types of spectral sparse dictionary $D = [D_1, D_2, \dots, D_K]$.

Step 1: Building a mixed spectral dictionary. Selecting m atoms randomly from each spectral dictionary D_i to constitute a mixed spectral dictionary D_0 .

Step 2: Preliminary reconstruction. The spectral spectrum D_0 is used to reconstruct the spectrum and the initial solution x_0 is obtained, which is the same as in the Algorithm 1.

Step 3: Calculate sparse representation coefficients of reconstructed spectrum x_0 for each type of sparse dictionary D_i .

$$a_i = \arg \min_{a_i} \|a_i\|_0 \quad s.t. \quad x_0 = D_i a_i, \quad i = 1, 2, \dots, K \quad (11.39)$$

Step 4: Calculate GINI sparsity for each sparse coefficient a_i .

$$G[i] = 1 - \frac{2}{l_1} \sum_{j=1}^N a_i[j] \frac{N - j + 1/2}{N} \quad (11.40)$$

Step 5: Category determination. Select the classification category c that corresponds to the largest indicator of GINI sparseness.

Step 6: Final reconstruction. Select dictionary D_c to reconstruct the object again and output the result, which is the same as Step 5 and 6 in Algorithm 1.

The common part of Algorithms 1 and 2 is to first reconstruct the target using a mixed dictionary extracted from various spectral dictionaries, then according to the preliminary reconstructed result, the category of the corresponding dictionary corresponding to the target is determined, and finally the spectrum sparse dictionary of the corresponding category is selected to reconstruct the target again. The difference is that the former classifies the preliminary reconstructed results by K-nearest neighbor algorithm based on known classification spectral training samples. The latter, by calculating the sparsity of the sparse representation of the initial reconstructed results under the spectral dictionary of each category, determines the classification and requires more computational effort.

Experiment 6 Choose Jarosite and Red Willow sample in the USGS_MIN and USGS_VEG spectral libraries to preliminarily reconstruct using a mixed spectral dictionary when the sampling rate is changed from 0.02 to 0.2 at 40 dB SNR and the signal to noise ratio is changed from 20 to 60 dB at 0.1 sample rate, and classify the preliminary reconstructed results using the above two strategies. Compare with the

correct classification results already calculated in Sect. 11.3.4 and calculate the accuracy of classification under 100 independent experiments. As shown in Figs. 11.17 and 11.18, when the sampling rate and signal-to-noise ratio are high, both algorithms can completely and accurately find the target corresponding category. However, the correct rate will decrease when the sampling rate and signal-to-noise ratio become lower, and the correct rate of Algorithm 1 is higher than that of Algorithm 2. It shows that the training sample information is best known when spectral reconstruction is performed based on classification dictionary under the condition that there is no a priori to the reconstructed object. Otherwise, it is not very reliable to rely on the sparse representation of the preliminary reconstruction result under the classification dictionary as classification criterion. However, in practice, all classification training

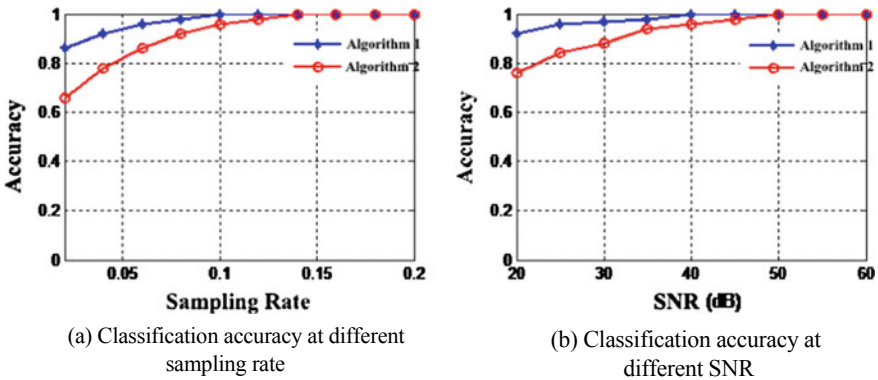


Fig. 11.17 The classification accuracy of Jarosite spectrum for initial reconstructed results at different sampling rates and signal-to-noise ratios

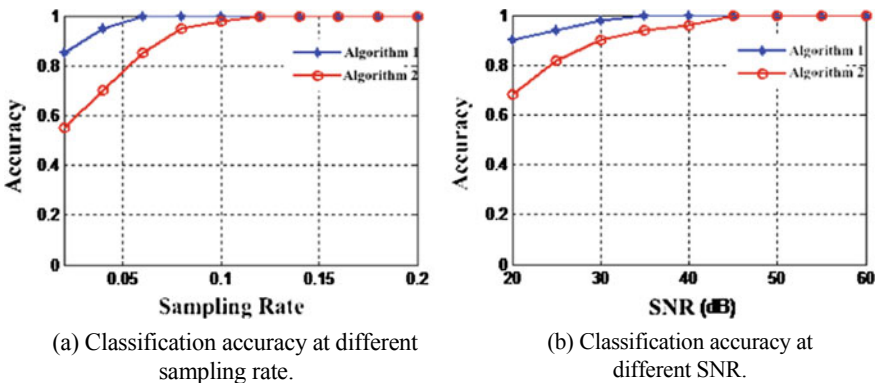


Fig. 11.18 The classification accuracy of Red Willow spectrum for initial reconstructed results at different sampling rates and signal-to-noise ratios

spectrum samples may not be saved. Under this condition, Algorithm 2 must be used to determine the dictionary category.

Experiment 7 To further verify the effect of adaptive selection strategies of spectral sparse dictionaries in hyperspectral correlation reconstruction, spectral data from the Indian Pines scene in the 0.4–2.5 μm wavelength range acquired by AVIRIS hyperspectral remote sensing imager is used for experiments. For hyperspectral data acquired by earth observation satellites, some spectral bands are greatly affected by atmospheric water vapor absorption. Several spectral bands affected by the atmosphere are removed, eventually 200 spectral bands of AVIRIS data are used in the experiments. In order to better compare, normalized preprocessing of all spectral data is applied. Figure 11.19 shows the spatial image of the 20th and 130th bands in the hyperspectral scene.

The selected spectral sample categories and sample sizes are shown in Table 11.1. To reduce the difference in DN values of images in different scenes, all spectral data are normalized. Spectral samples of each type in the sample database are trained in sparse dictionary according to the classification of natural attributes. The fast OMP algorithm with high efficiency is used as the sparse coding algorithm in the training process, and the KSVD algorithm is used as the dictionary update algorithm. In the training algorithm, the sparsity threshold L is set to 5, the number of iterations is 30, and each type of dictionary contains 100 atoms. Small dictionaries for different types of feature training in large dictionaries are listed in the order shown in Table 11.1.

Choose the spectral target of 128×128 spatial points outside the dictionary training area to perform correlation measurement and reconstruction experiment. Set the sampling rate to 0.1, the signal-to-noise ratio of the measured value to 40 dB, and the reconstruction algorithm is IRLS. Select different categories of sparse dictionaries in reconstruction based on target preferences in the unoptimized mixed dictionary

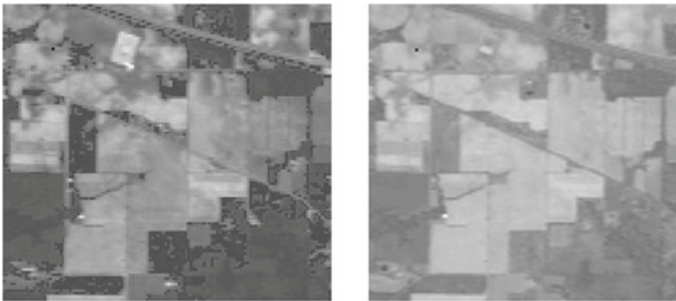
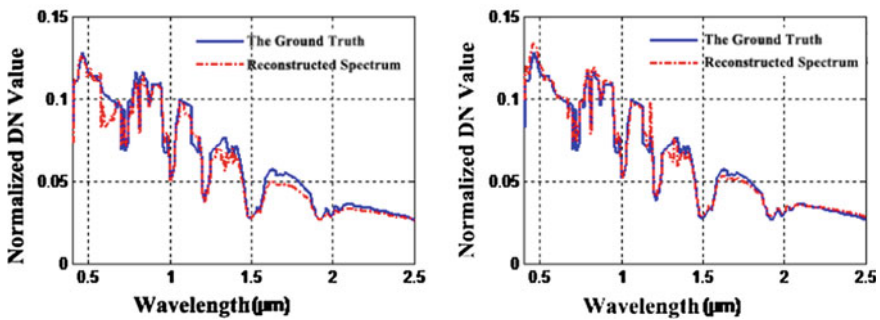


Fig. 11.19 AVIRIS hyperspectral scene in experiments (20th band and 130th band)

Table 11.1 The number of various types of features in the AVRIS training sample

Type	Tree	Corn	Wheat	Soybean
Number of samples	730	237	205	593

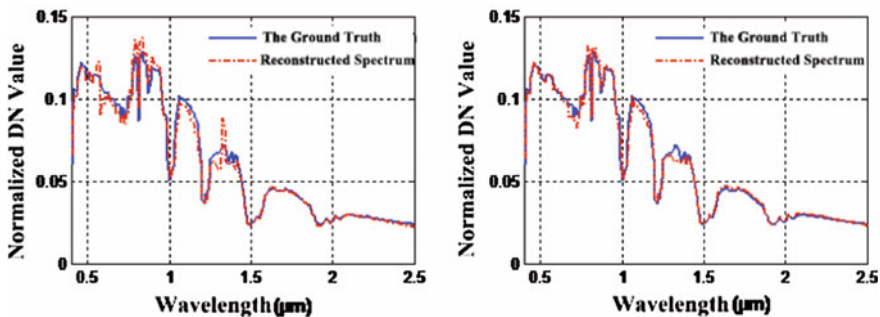
and Algorithms 1 and 2 respectively, comparing the results using different dictionary selection strategies. Figures 11.20 and 11.21 show the results of the correlation reconstruction of the corn spectrum and the alfalfa spectrum in the test scene. It can be seen that the spectral reconstruction accuracy of adaptive optimization after dictionary selection is higher than that of unoptimized spectral reconstruction accuracy. In the reconstruction of the two test spectral targets in this experiment, the selection results of the two dictionary selection strategies are consistent, so the results of the reconstruction are consistent. However, the results of the two strategies for dictionary adaptive selection of other spectral targets in the scene are not necessarily consistent. The overall RMSE and MSA errors were calculated for all reconstructed hyperspectral scenes. The results are shown in Table 11.2. It can be seen that the reconstruction accuracy of the dictionary adaptive selection method under the known spectral training samples is better than the reconstruction accuracy when the spectral training samples are unknown. However, no matter what kind of optimization method, the



(a) Before optimization (RMSE=2.13%).

(b) After optimization (RMSE=1.54%).

Fig. 11.20 The spectral reconstruction before and after the adaptive optimization of the dictionary selection for the corn spectrum in AVIRIS scene



(a) Before optimization (RMSE=1.77%).

(b) After optimization (RMSE=1.19%).

Fig. 11.21 The spectral reconstruction before and after the adaptive optimization of the dictionary selection for the alfalfa spectrum in AVIRIS scene

Table 11.2 Hyperspectral scene association reconstruction errors before and after different optimization strategy

	Before optimization	Algorithm 1	Algorithm 2
RMSE (%)	2.04	1.41	1.63
MSA	0.098	0.071	0.079

reconstruction accuracy has a certain degree of improvement compared to the non-adaptive selection of sparse dictionary.

11.4.2 *Optimized Adaptive Dictionary Selection Strategy Based on Spatially Neighboring Pixels*

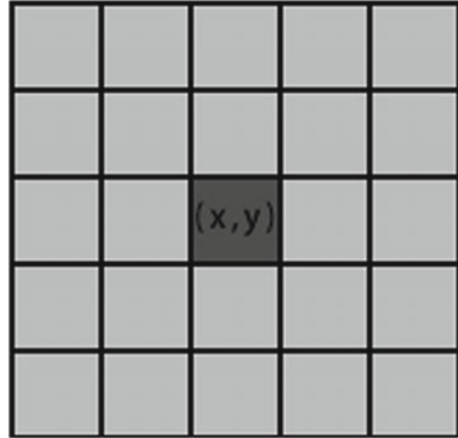
Usually the target of spatial proximity in a scene has a very high probability of being the same type. Figure 11.22 shows the real feature classification in the AVIRIS hyperspectral scene in the above experiment. In the figure, different types of objects are distinguished by different colors. It can be seen that the same type of objects has obvious spatial agglomeration. For this feature of hyperspectral features scenes, it is very likely that the same type of spectral sparse dictionary should be used to reconstruct the spatially adjacent spectral targets in the correlation reconstruction. Therefore, for the correlation reconstruction of hyperspectral scenes containing spatial information, the similarity of neighboring positions should be considered in the dictionary type adaptive selection algorithm in the reconstruction.

In this article, the neighborhood of a spatial point is defined as the point within the range of up, down, left, and right by two rows and two columns, as shown in Fig. 11.23. The point of the edge location is beyond the point of interest outside the scene range. The space point in this scene has a maximum of 24 neighborhoods

Fig. 11.22 Real feature classification of the AVIRIS target scene



Fig. 11.23 Point (x, y) and its neighborhood



and a minimum of 8 neighborhoods. The reason why the simpler 4-neighborhood or 8-neighborhood model is not adopted is because the number of neighboring sample points is too small (especially at the edge position) may affect the accuracy of classification judgment, and instead of using larger neighborhoods is to avoid excessive calculations.

The adaptive dictionary selection strategy that considers the spatial domain similarity is: Firstly, Algorithm 1 or Algorithm 2 is used to select the corresponding type of spectral sparse dictionary for each spatial point in the target scene.

Assuming a common K -type spectral dictionary, calculate the index for each spatial point x in the scene.

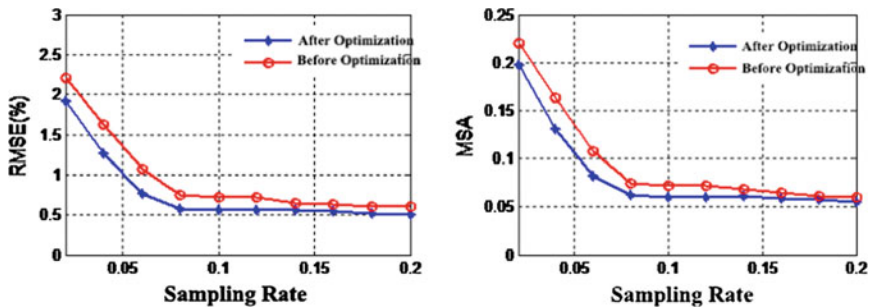
$$P[i] = \sum_{j=1}^n \frac{1}{d_j} I_{j,i}, i = 1, 2, \dots, K \tag{11.41}$$

where n is the number of neighborhood points of x and d_j is the spatial distance from the neighborhood point x_j to point x . The function $I_{j,i}$ is

$$I_{j,i} = \begin{cases} 1 & \text{if } x_j \in \text{class}\{i\} \\ 0 & \text{else} \end{cases} \tag{11.42}$$

The meaning of the index $P[i]$ is the total number of distance-weighted targets of the i -th category in the neighborhood of point x . The corresponding dictionary type of point x is updated to the serial number of the largest value in P . If the updated dictionary is different from the original dictionary, the target spectrum of the position is reconstructed again using the updated dictionary.

Experiment 8 The adaptive selection of sparse dictionaries in the reconstruction in Experiment 7 is modified using the neighborhood similarity feature. Figure 11.24 shows the variation of RMSE and MSA error with sampling rate for adaptive recon-



(a) RMSE error at different sampling rate. (b) MSA error at different sampling rate.

Fig. 11.24 The reconstruction error with different sampling rate under adaptive reconstruction of neighborhood optimization

struction before and after the neighborhood optimization. It can be seen that the reconstruction accuracy of the target scene is improved after the neighborhood optimization method. It shows that the strategy of adopting neighborhood optimization can correct errors in dictionary selection during reconstruction, and it is an effective method for optimizing dictionary selection in hyperspectral correlation reconstruction.

11.5 Application of Hyper-spectral Sparse Dictionary in Real Imaging System

The former chapters discussed the construction of correlation hyperspectral remote sensing sparse dictionaries and their optimization in the reconstruction. However, for different models of actual hyperspectral imagery systems, how to make the corresponding modulation and optimization of the algorithm in the reconstruction based on hyperspectral sparse dictionary after adding spatial information. Whether the construction of hyperspectral sparse dictionaries validated in simulation experiments is still valid in real systems is a question that must be answered in the theoretical application of this paper. Therefore, this chapter will focus on two real imaging systems, the hyperspectral correlation imaging system based on LCTF and the single-exposure hyperspectral correlation imaging systems based on phase modulation, the spatial-spectral joint correlation reconstruction method under hyperspectral sparse dictionaries is proposed and the reconstruction of actual imaging target scene is tested.

11.5.1 Sparse Dictionary Application in the Hyperspectral Correlation Imaging System Based on LCTF

A principle demonstration system is built on the laboratory optical platform based on the correlation imaging model constructed above, as shown in Fig. 11.25. The system includes halogen illumination source, collimating lens, Digital Micromirror Device (DMD), imaging target, Liquid Crystal Tunable Filter (LCTF) and single point detector. The spectrum of LCTF in the experiment is in the visible spectrum range of 400–720 nm. The correlation measurement process is: After the collimated light source illuminates the DMD, the spatial coding on the DMD is projected onto the imaging target. The spatially-encoded reflected light of the imaging target is then spectrally encoded through the LCTF, and finally the spatially and spectrally-coordinated light field converges to the detector to complete a correlation measurement. Then the spatial and spectral encodings are updated synchronously by the control of the computer program and the next correlation measurement is performed. After all the measurements are completed, the measurement values as well as the spatial and spectral coding matrix and the constructed hyperspectral sparse dictionary is processed using the above method to perform correlation reconstruction.

The corresponding measurement matrix for spatial coding using DMD is $(0, 1)$ random matrix. In correlation reconstruction $(-1, 1)$ random matrix is a more ideal measurement matrix. Therefore, differential preprocessing on the measurement data and measurement matrix before reconstruction is performed.

$$y'[i] = y[i] - y[i + 1], i = 1, 2, \dots, m \quad (11.43)$$

$$\Phi[i, j] = \Phi[i, j] - \Phi[i + 1, j], i = 1, 2, \dots, m; j = 1, 2, \dots, nb \quad (11.44)$$

Fig. 11.25 The hyperspectral correlation imaging demonstration system based on LCTF

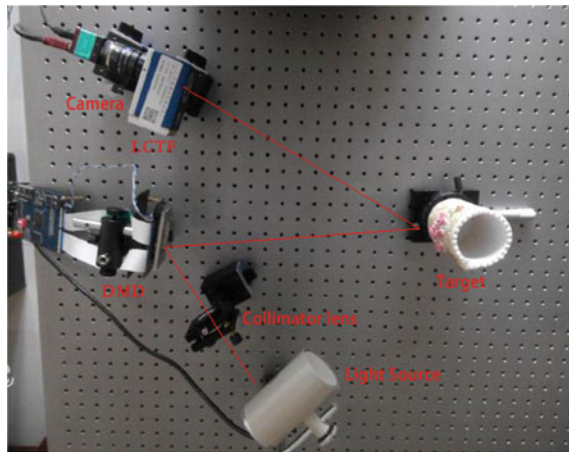


Image a square area in the color bookmark shown in Fig. 11.26 and perform correlation reconstruction based on the model above. The sampling rate in the experiment is 0.5. The reconstructed hyperspectral image size is $80 \times 80 \times 28$. Figure 11.27 shows the reconstructed target images of different spectral bands in the 550–720 nm wavelength range. Using 450, 550 and 700 nm bands of the reconstructed images for blue, green and red component respectively to synthesize a colored image. The color composite image is shown in Fig. 11.28. Figure 11.29 shows the spectral curves of the red points (coordinates (65,20)), green points (coordinates (42,27)) and white points (coordinates (20, 42)) extracted from the reconstructed image. It can be seen that the reconstructed target image color is not much different from the real scene, spatial information reconstruction is accurate, and no obvious distortion and distortion appear.

However, due to the long measurement and reconstruction time required for hyperspectral correlation imaging, the number of spatial pixels in the reconstructed image is not set sufficiently in the experiment. Therefore, some spatial details in the real scene cannot be reflected in the reconstructed image. However, the experimental verification of the model and algorithm in this paper is enough.

Set different sampling rates in the experiment, as shown in Fig. 11.30. It can be seen that when the sampling rate is low, the reconstructed image will appear blurred, and the spatial information will be greatly degraded. The color information does not change much, indicating that the reconstructed model proposed in this paper has excellent performance in spectral reconstruction.

As a comparison, only the traditional spatial association model is used to reconstruct the target scene under the same experimental conditions, as shown in Fig. 11.31. It can be seen that the image obtained by using only the spatial correlation reconstruction model under low sampling not only has obvious degradation of spatial information, but also has distorted image color, i.e., the spectral reconstruction is

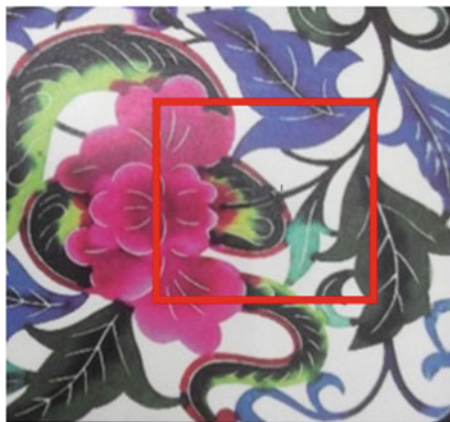


Fig. 11.26 The imaging target in the experiment



Fig. 11.27 The reconstructed images of different spectral bands

Fig. 11.28 The synthetic colored reconstructed images

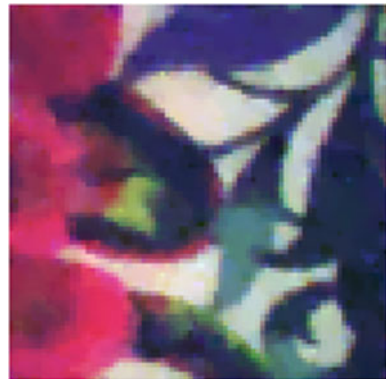
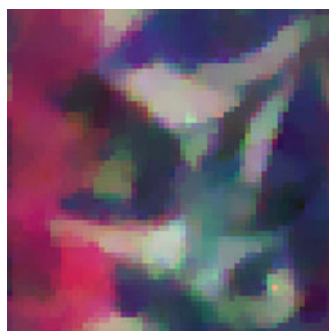
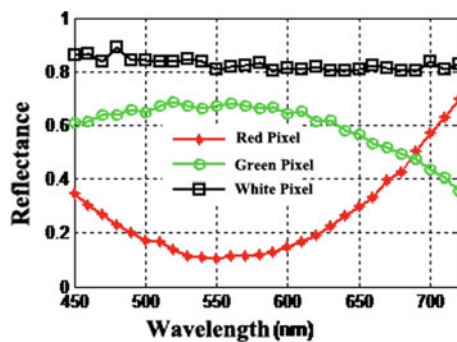
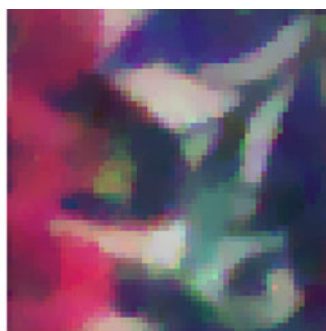


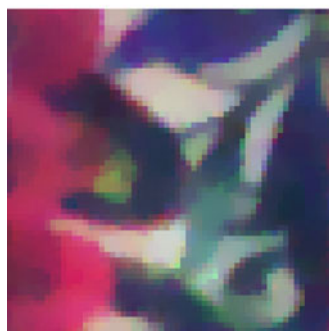
Fig. 11.29 The reconstruct spectral curves at different spatial positions



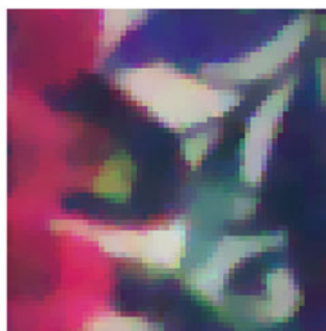
(a) Sampling rate=0.15.



(b) Sampling rate=0.2.



(c) Sampling rate=0.25.



(d) Sampling rate=0.3.

Fig. 11.30 The reconstructed color image at different sampling rate

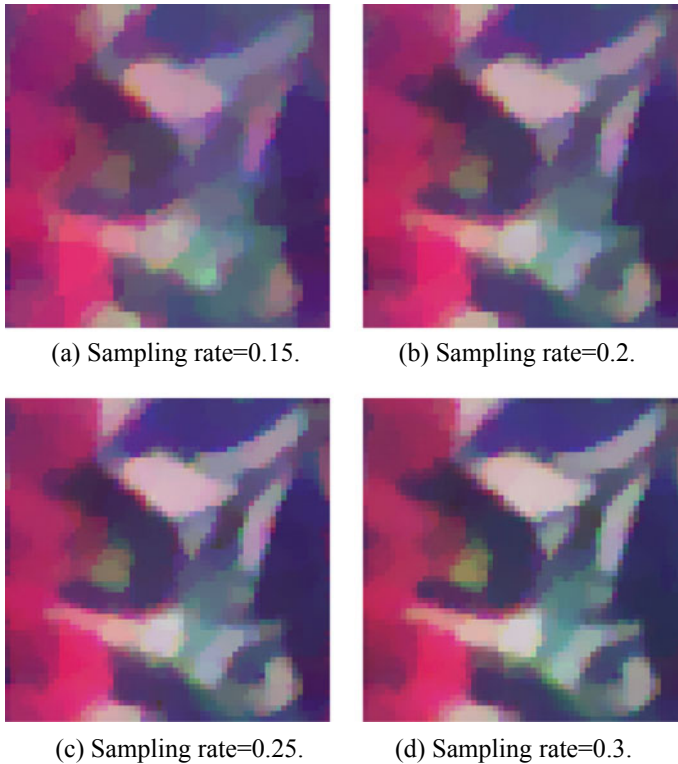


Fig. 11.31 The reconstructed color image using only spatial correlation model at different sampling rate

not accurate. The validity of the method of adding spectral sparse dictionary into the reconstruction is verified in the comparison experiment.

11.5.2 Sparse Dictionary Application in the Single Exposure Correlation Hyperspectral Imaging System

Single exposure correlation hyperspectral imaging system based on phase modulation has the advantages of high measurement efficiency, high energy utilization, and high resolution [33]. It is the world's first correlation imaging system that successfully achieves the observation of earth observations on an aviation platform. The system consists of three modules: pre-imaging module, modulation detection module and demodulation reconstruction module. The pre-imaging module includes an imaging objective lens and a wide-band filter, and its function is to focus and project the distant imaging scene on the first image plane for subsequent processing. The modulation

detection module is the core module of the system and includes a random phase modulator and a photodetection array. The random phase modulator forms a speckle field using light diffraction principle, enabling a spatial-spectral mixed coding of the target scene. The demodulation and reconstruction module perform correlation reconstruction through the received measurement data of the photodetection array and the previously measured random phase modulation mode to obtain the final scene image.

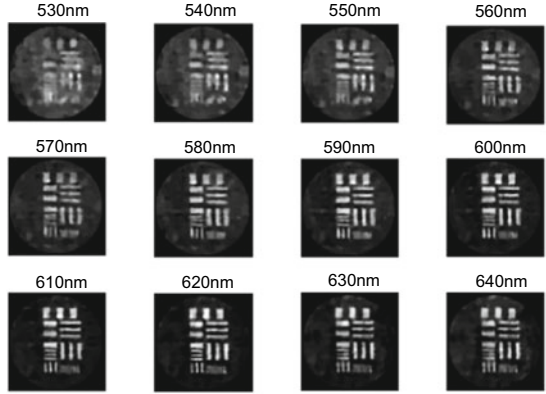
The single exposure hyperspectral correlation imaging system based on phase modulation was mounted on a high air balloon to perform remote sensing imaging experiments on the scene near Beijing Bei'anhe Airport. The imaging scene includes natural scenes such as vehicles, stadiums, and houses, as well as artificial scenes such as three-line targets, gray-scale targets, and color targets. Using the reconstruction method based on spatial-spectral sparse model proposed above, the experimental data are subjected to hyperspectral correlation reconstruction and compared with other methods (Fig. 11.32).

Computation of correlation measurement data for three-line target scenes that characterize spatial resolution is executed. Figure 11.33 shows the reconstructed three-line target image at different spectral bands using the TV algorithm [34], TVSS + NU algorithm [35] and the spatial-spectral sparse model reconstruction method proposed in this paper. As can be seen from the visual effects in the figure, due to the lack of constraints on the spectrum, the reconstructed scenes by TV algorithm have different extent of noise in different spectral regions, and the image quality is unstable. For example, the image quality at 530 nm wavelength is significantly worse than other spectral images. The target scene reconstructed by the TVSS + NU algorithm combined with kernel norm constraint and reference spectral segment optimization method is clearer, and the image quality of different spectral segments is relatively stable. The scene reconstructed by the proposed algorithm is also clear

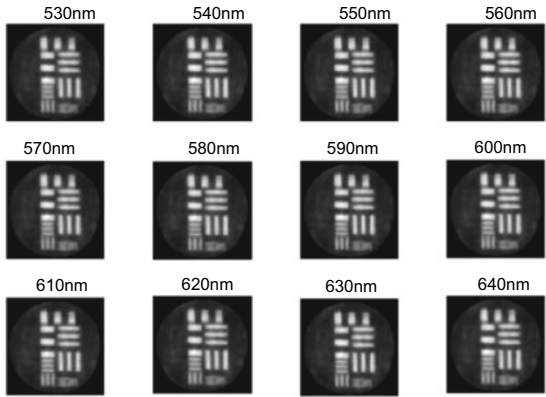


Fig. 11.32 The high air ball platform with the correlation imaging system

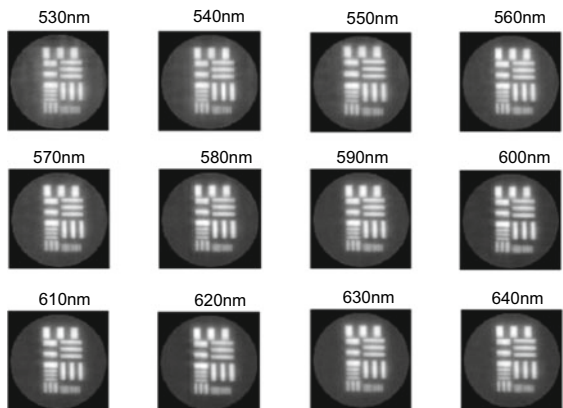
Fig. 11.33 The reconstructed three-line target hyperspectral image by different algorithms



(a) TV algorithm.



(b) TVSS+NU algorithm.



(c) The proposed algorithm.

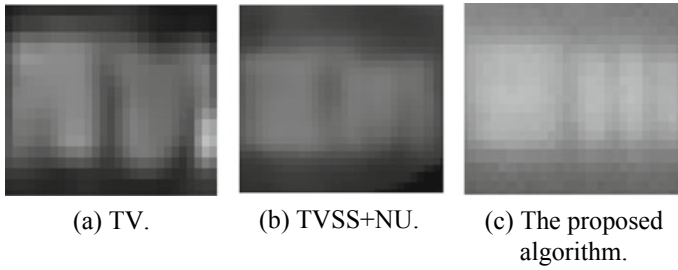


Fig. 11.34 Enlarged part of the reconstructed images of three-line target images with different algorithms

and the image quality is stable. At the same time, due to the use of sparse dictionaries to correct the reconstructed image, the noise of the image is lower than that of the TVSS + NU algorithm.

Examine the resolution of the three-line target in the reconstructed image. It is easy to see that the three algorithms can distinguish groups of three-line stripe with large spacing, and the stripes with the highest resolution in the lower right corner of the image is more blurred. The right corner of the image at the 640 nm wavelength reconstructed by the three algorithms is enlarged to observe, as shown in Fig. 11.34. The left lateral stripes cannot be distinguished in the reconstructed images by the three algorithms. For the right vertical stripes, the rightmost one can be barely distinguished in the TV algorithm reconstructed image and cannot be distinguished in the TVSS + NU algorithm reconstructed image, while all the strips can be distinguished in the images reconstructed by the proposed method. The spatial resolution of the image reconstructed by the proposed algorithm is better than other methods.

11.6 Conclusions

This article is aimed at the problem that the correlation hyperspectral remote sensing imaging system is not ideal for sparse representation of complex ground objects in image reconstruction, resulting in low quality of reconstructed images and spectrum. By combining the sparse coding and dictionary learning theory of signal processing with correlation remote sensing imaging, the related theories and methods for the construction of correlation hyperspectral remote sensing sparse dictionaries are studied. Through the construction of hyperspectral remote sensing sparse dictionaries, optimization in reconstruction, and application in actual imaging systems, a set of sparse dictionaries for correlation hyperspectral imaging is constructed. Compared with existing methods such as spatial total variational constraints and hyperspectral image kernel norm constraints, the ability to sparsify hyperspectral remote sensing scenes and the reconstruction quality of hyperspectral images has been effectively

improved. It has guiding significance for the development of the correlation imaging field. The main research contents and conclusions are as follows:

- (1) A general model of correlation hyperspectral measurement and reconstruction based on spectral sparsity is deduced. For the dictionary construction model based on training algorithm, to solve the problem that the number of available spectral library samples is limited and it is difficult to carry out effective dictionary training, a method of spectral sample expansion by various transformations such as stretching, shifting, and interception is proposed. A variety of spectral sparse dictionaries are used in spectral correlation reconstruction and comparative analysis is performed using a variety of dictionary update methods such as MOD, OLD, and KSVD. Aiming at the problem of sparse representation of the spectrum of hybrid dictionary, this paper proposes the idea of classification sparse dictionary and uses the K-means clustering method to classify non-supervised training samples.
- (2) An adaptive optimization method for correlation reconstruction based on hyperspectral sparse dictionary is studied. For the problem that it is difficult to select a suitable type of spectral sparse dictionary for the target without prior information in the reconstructing reconstruction, the adaptive selection strategy of the multi-step spectral sparse dictionary based on the K-nearest neighbor classification method and sparsity optimization method is proposed. Further correction of dictionary selection strategies based on the similarity of neighboring pixels in hyperspectral scenes is proposed to solve the problem of sparse dictionary selection without prior information. Experiments show that the accuracy of the selection of different types of dictionaries can reach more than 95% at the 5% sampling rate, and the accuracy of the hyperspectral correlation reconstruction can be improved by more than 20%.
- (3) The reconstruction and optimization methods of the spatial-spectral joint reconstruction of two types of the correlation hyperspectral imaging systems based on the sparse dictionary are studied: the hyperspectral correlation imaging system based on LCTF and the single-exposure hyperspectral correlation imaging systems based on phase modulation. In comparison with other reconstruction methods, the proposed method shows better performance in both spatial image quality and spectral accuracy.

Acknowledgements This work is supported in part by the National High Technology Research and Development Program of China under Grant 2014AA123201 and the National Key Research and Development Program of China under Grant 2016YFB0500402.

References

1. T.B. Pittman et al., Optical imaging by means of two-photon quantum entanglement. *Phys. Rev. A* **52**(5), R3429–R3432 (1995)
2. A. Gatti et al., Coherent imaging with pseudo-thermal incoherent light. *J. Mod. Opt.* **53**(5-6), 739–760 (2006)
3. Y. Yan, H. Dai, X. Liu, W. He, Q. Chen, G. Gu, Colored adaptive compressed imaging with a single photodiode. *Appl. Opt.* **55**(14), 3711 (2016)
4. R. Boyd, Promises and Challenges of Ghost Imaging, in *Signal Recovery and Synthesis* (2011)
5. W. Gong, S. Han, Super-resolution ghost imaging via compressive sampling reconstruction. *Physics* (2009)
6. R.E. Meyers, K.S. Deacon, Y. Shih, Turbulence-free ghost imaging. *Appl. Phys. Lett.* **98**(11), 111115–111115-3 (2011)
7. S. Yuan, X. Liu, X. Zhou, et al., Multiple-object ghost imaging with a single-pixel detector. *J. Opt.* **1–7** (2015)
8. Clemente P, Durán V, Torrescompany V, et al., Optical encryption based on computational ghost imaging. *Opt. Lett.* **35**(14), 2391–2393 (2010)
9. J.M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, J. Chanussot, Hyperspectral remote sensing data analysis and future challenges. *IEEE Geosci. Remote Sens. Mag.* **1**(2), 6–36 (2013)
10. G. Brida, M.V. Chekhova, G.A. Fornaro, M. Genovese, E.D. Lopaeva, I.R. Berchera, Systematic analysis of signal-to-noise ratio in bipartite ghost imaging with classical and quantum light. *Phys. Rev. A* **83**(6), 63807 (2011)
11. V.K.J. Astola, Compressive sensing computational ghost imaging. *J. Opt. Soc. Am. A:* **29**(8), 1556–1567 (2012)
12. E.J. Candes, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **52**(2), 489–509 (2006)
13. Y. Song, The application of compressed sensing algorithm based on total variation method into ghost image reconstruction, in *International Conference on Optoelectronics and Microelectronics Technology and Application*, p. 102440X (2017)
14. L. Zhang, Y. Zhang, W. Wei, 3D total variation hyperspectral compressive sensing using unmixing. *Geosci. Remote Sens. Sympos. (IGARSS)* **2014**, 2961–2964 (2014)
15. F. Yan, C.Y. JiaYingbiao et al., Compressed sensing projection and compound regularizer reconstruction for hyperspectral images. *Acta Aeronautica et Astronautica Sinica* **33**(8), 1466–1473 (2012)
16. M. Golbabaee, P. Vandergheynst, Compressed sensing of simultaneous low-rank and joint-sparse matrices. *IEEE Transac. Inf. Theor.* (2012)
17. F.Y. JiaYingbiao, W. Zhongliang et al., Hyperspectral compressive sensing recovery via spectrum structure similarity. *J. Electron. Inf. Technol.* **6**, 1406–1412 (2014)
18. K. Kreutz-Delgado, J.F. Murray, B.D. Rao et al., Dictionary learning algorithms for sparse representation. *Neural Comput.* **15**(2), 349–396 (2014)
19. M.F. Duarte, M.A. Davenport, M.B. Wakin, J.N. Laska, D. Takhar, K.F. Kelly, et al., Multiscale random projections for compressive classification. *IEEE Int. Conf. Image Process.* **6**, VI-161–VI-164
20. J. Wright, A. Ganesh, Z. Zhou, A. Wagner, Y. Ma, Demo: Robust face recognition via sparse representation. *IEEE Int. Conf. Autom. Face Gesture Recog.* **31**, 1–2 (2009)
21. L.W. Kang, C.Y. Hsu, H.W. Chen, C.S. Lu, Secure SIFT-based sparse representation for image copy detection and recognition. *IEEE Int. Conf. Multimedia Expo*, 1248–1253 (2010)
22. K. Estabridis, Automatic target recognition via sparse representations, **7696**(6), 701–712 (2010)
23. Q. Wang, H. Xu, L. Ma, et al., Hyperspectral compressive sensing imaging via spectral sparse constraint. In: *International Conference on Photonics, Optics and Laser Technology* (2018)
24. S.G. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.* **41**(12), 3397–3415 (1993)

25. J. Tropp, A.C. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory* **53**(12), 4655–4666 (2007)
26. D. Needell, R. Vershynin, Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Found. Comput. Math.* **9**(3), 317–334 (2009)
27. D.L. Donoho, Y. Tsaig, I. Drori, J.L. Starck, Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. *IEEE Trans. Inf. Theory* **58**(2), 1094–1121 (2012)
28. J. Deng, G. Ren, Y. Jin, W. Ning, Iterative weighted gradient projection for sparse reconstruction. *Inf. Technol. J.* **10**(7), 1409–1414 (2011)
29. H. Mohimani, M. Babaie-Zadeh, C. Jutten, A fast approach for overcomplete sparse decomposition based on smoothed l_0 norm. *IEEE Press.* (2009)
30. P. Yang, F. Yan, F. Yang, Sparse array synthesis with regularized focuss algorithm, vol. 9, no. 6, pp. 1406–1407 (2013)
31. M.J. Lai, Y. Xu, W. Yin, Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization. *Siam J. Numer. Anal.* **51**(2), 927–957 (2013)
32. R. Rubinstein, T. Peleg, M. Elad, Analysis k-svd: a dictionary-learning algorithm for the analysis sparse model. *IEEE Trans. Signal Process.* **61**(3), 661–677 (2013)
33. J. Wu, S. Xia, Y. Hong, C. Zhe, Z. Liu, S. Tan et al., Snapshot compressive imaging by phase modulation. *Acta Optica Sinica* **34**(10), 113–120 (2014)
34. C. Li, An efficient algorithm for total variation regularization with applications to the single pixel camera and compressive sensing. *Dissertations & Theses–Gradworks* (2010)
35. S. Tan, Z. Liu, E. Li, S. Han, Hyperspectral compressed sensing based on prior images constrained. *Acta Optica Sinica* **35**(8), 112–120 (2015)

Author Index

A

Achten, F. J., 57
Akbar, Demiral, 221
Altan, Hakan, 221
Amaral, Gustavo, 33

B

Bierl, Rudolf, 169
Bikorimana, Simeon, 119

C

Calliari, Felipe, 33
Carrasco-Ortiz, Sergio L., 197
Chang, H. S., 143
Chen, H. S., 143
Chou, Emin, 143

D

Dorsinville, Roger, 119

H

Hausler, Peter, 169
Herrera, Luis, 33
Hossain, Abdullah, 119
Huang, Jack Jia-Sheng, 143

J

Jan, Yu-Heng, 143

K

Kusoglu-Sarikaya, Cemre, 221

L

Lee, S. K., 143
Li, Chuanrong, 239
Llorente, Roberto, 197

M

Ma, Lingling, 239
Molin, D., 57
Morant, Maria, 197

N

Nahear, Rotem, 93
Ni, C. J., 143
Noach, Salman, 93

R

Roth, Carina, 169

S

Schranz, Ágoston, 75
Sheintop, Uzziel, 93
Shi, Jin-Wei, 143
Soãres, Schubert, 1

T

Tang, Lingli, 239

UUdvary, Eszter, [75](#)Ummay, Muhammad, [119](#)**V**Valero, Eduardo, [197](#)Vitzthumecker, Thomas, [169](#)Von der Weid, Jean Pierre, [33](#)**W**Wang, Qi, [239](#)Wang, Xinhong, [239](#)**X**Xu, Hong, [239](#)**Z**Zhou, Yongsheng, [239](#)

Subject Index

A

Absorption, 128
Air concentration, 199, 204, 212, 218

B

Beam combination, 122
Beat signal, 122
Bubble, 197–209, 211–215, 217, 218

C

Cherenkov, 127
CMOS, 197, 200–202, 204, 211, 217
Coherence length, 125
Compound-ring, 129
Continuous wave, 121
Correlation imaging, 240–244, 249, 251, 252, 274, 278, 279, 282
Correlation reconstruction, 242, 243, 251, 263–265, 270, 271, 273–275, 282

D

Defects, 2–7, 9, 10, 13, 15, 24–27
Dictionary learning, 243, 249, 250, 262, 281
Difference frequency mixing, 122
Dual wavelength, 130

E

Excitation via tilted optical pump beams, 127

F

Fiber, 129
Fiber hybrid, 129
FPGA, 39–41, 45

G

Geometry, 127
Glow discharge detectors, 222

H

Hyperspectral correlation imaging, 243, 273–275, 279, 282
Hyperspectral correlation reconstruction, 261, 269, 273, 279, 282
Hyperspectral imaging, 241
Hyperspectral reconstruction, 242, 243
Hyperspectral remote sensing, 241, 261, 273, 281
Hyperspectral remote sensing correlation imaging, 251

I

Ionizing, 121

L

Laser, 129
Laser triangulation, 2, 14, 30

M

Micro-Opto-Electro-Mechanical System (MOEMS), 169
 Miniaturized SPR devices, 169
 MIR lasers, 93
 Mm-waves, 221
 Multiphase flow, 197–200, 211, 218
 Multiphase water flow, 197, 217

N

Near infrared lasers, 121
 Noise error, 20, 24
 Non-linear, 128
 Non-linear effects, 125
 Non-resonant optical rectification, 126

O

Optical pumping, 121
 Optical scatter, 8, 13, 15, 30
 Optical surface, 15, 22, 26

P

Parametric oscillation, 121
 Passive phase-locking, 129
 Periodically inverted crystals, 121
 Phase matching, 125
 Photoconductive switches, 121
 Photo mixers, 121
 Photon-counting Optical Time Domain Reflectometer (OTDR), 34, 38–41, 44–46, 53, 54
 Photon-Counting Optical Time Domain Reflectometry, 33
 Photonic sensor, 197, 199, 200, 217
 PIC/MCC, 221
 Post-processing algorithms, 213, 218
 Processing algorithms, 197, 199, 211

R

Radiation, 121
 Reconstruction algorithm, 240, 241, 251, 252, 254, 255, 257, 262, 269
 Resonant optical rectification, 124

S

Semiconductor optical amplifier, 129
 Semiconductor wafers, 1, 6, 9, 15, 25
 Signal processing, 38, 42, 43, 50
 SLM, 129
 Solid-state lasers, 93
 Sparse representation, 243–247, 249, 252, 256, 267, 268, 281, 282
 Spectral correlation reconstruction, 253, 254, 262–264, 282
 Spectral reconstruction, 254, 268, 270, 275
 Speed, 197, 199, 204, 209, 210, 213, 214, 216, 217
 Surface Plasmon Resonance imaging (SPRi), 169

T

Terahertz (THz), 119, 121, 221
 Thin films, 1, 3, 13, 29
 THz attenuation, 126
 Tunable, 129
 Tunable laser, 94, 101, 104, 105
 Tunable pulsed laser, 94, 111
 Tunable solid-state lasers, 101
 Tunable Tm based lasers, 93
 Tunable Tm lasers, 113

W

Water flow, 217