Polina Golland Nobuhiko Hata
Christian Barillot Joachim Hornegger
Robert Howe (Eds.)

LNCS 8675

Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014

17th International Conference
Boston, MA, USA, September 14–18, 2014
Proceedings, Part III

Springer

Lecture Notes in Computer Science 8675

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Polina Golland Nobuhiko Hata
Christian Barillot Joachim Hornegger
Robert Howe (Eds.)

Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014

17th International Conference
Boston, MA, USA, September 14-18, 2014
Proceedings, Part III

Springer

Volume Editors

Polina Golland
Massachusetts Institute of Technology
Cambridge, MA, USA
E-mail: polina@csail.mit.edu

Nobuhiko Hata
Brigham and Women's Hospital
Harvard Medical School, Boston, MA, USA
E-mail: hata@bwh.harvard.edu

Christian Barillot
IRISA, Rennes, France
E-mail: christian.barillot@irisa.fr

Joachim Hornegger
Friedrich-Alexander University Erlangen-Nuremberg
Erlangen, Germany
E-mail: joachim.hornegger@fau.de

Robert Howe
Harvard University, Cambridge, MA, USA
E-mail: howe@seas.harvard.edu

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-319-10442-3 e-ISBN 978-3-319-10443-0
DOI 10.1007/978-3-319-10443-0
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014946218

LNCS Sublibrary: SL 6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The 17th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2014) was held in Boston, USA, at the Massachusetts Institute of Technology (MIT) and Harvard Medical School during September 14-18, 2014. We were delighted to welcome the conference back to the location of the very first MICCAI meeting that took place on MIT campus in 1998. Over the last 16 years, the MICCAI conferences have become a premier international event, with papers of high standard addressing open problems in the multidisciplinary fields of biomedical image computing, computer-assisted intervention, and medical robotics. The conference attracts leading scientists, engineers, and clinicians from a wide range of disciplines.

This year, we received a record number of 862 submissions. These covered medical image computing (functional and diffusion image analysis, segmentation, physical and functional modeling, shape analysis, atlases and statistical models, registration, data fusion and multiscale analysis), computer-assisted interventions and robotics (planning and image guidance of interventions, simulation and training systems, clinical platforms, visualization and feedback, robotics and human–robot interaction), and clinical imaging and biomarkers (computer-aided diagnosis, organ/system specific applications, molecular and optical imaging and imaging biomarkers). A careful systematic review process was carried out to create the most exciting scientific program for MICCAI 2014. The Program Committee (PC) of the conference was composed of 52 experts recognized internationally in the main topics of the conference. Each submission was assigned to a primary PC member who recruited between three and four external reviewers for each paper based on their expertise and the topic of the paper. The external reviewers provided double-blind reviews of the papers. Each submission without consensus among the external reviewers was assigned to two secondary PC members and was invited to submit a rebuttal followed by discussion among the external reviewers. Each secondary PC member made recommendations to the PC while taking into account the external reviews, the rebuttal, and the discussion. The list of accepted papers was finalized during a two-day PC meeting held at MIT during May 17-18, 2014, based on the scores and rankings provided by the PC members and external reviewers and on the discussion among the PC members. In all, we accepted 253 papers (29%) to be included in the proceedings of MICCAI 2014 and presented as posters during the meeting. Of these, 36 were selected for podium presentation (4%). We congratulate those who had papers accepted and encourage those who did not to persevere and submit again next year. Selection of papers for MICCAI is a competitive process and with such a strong submission pool it is inevitable that many good papers could not be included in the final program. We sympathize with the authors whose papers were rejected; we had our own share of rejected papers this year!

In addition to the main conference, MICCAI 2014 offered a rich program of workshops, computational challenges, and tutorials. We received a fantastic set of proposals that resulted in an exciting, diverse, and high-quality program. The workshops provided a comprehensive coverage of topics not fully explored during the main conference and of emerging areas of MICCAI; the computational challenges explored empirical solutions to hard open problems; the tutorials provided educational material for training new professionals in the field. We are grateful to all workshop, challenge, and tutorial organizers for making these events a success and to the workshop chairs for creating such a great program.

MICCAI 2014 introduced a completely new Educational Challenge, conceived, organized, and run by the MICCAI Student Board. The long-term goal is to create a video library of educational presentations for students entering the fields. The Educational Challenge was a great step in that direction, and we hope MICCAI will continue to support this effort. Our many thanks go out to the students who organized the challenge. We would also like to thank our invited speaker Neville Hogan (MIT, USA) for his presentation on the use of robots for rehabilitation.

We thank the external reviewers and the PC for volunteering their time and judgement to provide high-quality reviews and ensure a fair paper selection process. The continued improvement in the quality of the conference depends entirely on this tremendous effort. We thank James Stewart of *precisionconference.com* for the efficient organization of the website and amazingly fast responses to our questions and requests for changes. The conference would not be possible without the commitment and hard work of the MIT Conference Services staff that contributed tremendous amount of effort and energy to make sure all the logistics of the meeting ran smoothly. Our special thanks go out to Amy Hansen, who singlehandedly compiled the conference proceedings and conference program brochures, spent many hours in communications with the authors to ensure their papers are properly included in the proceedings, and handling many other aspects of the paper submission process. We also thank all the session chairs for managing and coordinating the presentations during the conference.

We thank the MICCAI Society for providing valuable input and support for the conference. We were delighted to have had a chance to organize a 10th anniversary celebration for the society. Many happy returns! Last but not least, we would like to thank all our sponsors for their kind support. Their generosity ensured the highest quality of the conference and essential support to students and young researchers.

It was our pleasure to welcome MICCAI 2014 participants to Boston. We look forward to seeing you all again next year in Munich, Germany!

September 2014

Polina Golland
Nobuhiko Hata
Christian Barillot
Joachim Hornegger
Robert Howe

Organization

General Chairs

Polina Golland	Massachusetts Institute of Technology, USA
Nobuhiko Hata	Harvard Medical School, USA

Program Chairs

Polina Golland	Massachusetts Institute of Technology, USA
Christian Barrilot	IRISA, France
Joachim Hornegger	Friedrich Alexander University, Erlangen-Nuremberg, Germany
Robert Howe	Harvard Medical School, USA

Workshop/Tutorial/Challenge Chairs

Georg Langs	Medical University of Vienna, Austria
Mehdi Moradi	University of British Columbia, Canada
Sonia Pujol	Harvard Medical School, USA
Martin Styner	University of Norther Carolina Chapel Hill, USA

Local Arrangements

MIT Conference Services

Student Liaison

Adrian Dalca	MIT, Cambridge, MA, USA

Program Committee

Purang Abolmaesumi	University of British Columbia, Canada
Stephen Aylward	Kitware, USA
Dean Barratt	University College London, UK
Dorin Comaniciu	Siemens, USA
Christos Davatzikos	University of Pennsylvania, USA
Marleen de Bruijne	Erasmus MC Rotterdam, The Netherlands
James Duncan	Yale University, USA
Randy Ellis	Queen's University, Canada
Gabor Fichtinger	Queen's University, Canada

James Gee	University of Pennsylvania, USA
Guido Gerig	University of Utah, USA
Miguel A. Gonzalez Ballester	ICREA-Universitat Pompeu Fabra, Spain
Leo Grady	HeartFlow, USA
Havit Greenspan	Tel Aviv University, Israel
Gregory D. Hager	Johns Hopkins University, USA
Ghassan Hamameh	Simon Fraser University, Canada
Ameet Jain	Philips, USA
Ron Kikinis	Harvard Medical School, USA
Ender Konukoglu	Harvard Medical School, USA
Hongen Liao	Tsinghua University, China
Huafeng Liu	Zhejiang University, China
Dimitris Metaxas	Rutgers University, USA
Mehdi Moradi	University of British Columbia, Canada
Kensaku Mori	Nagoya University, Japan
Mads Nielsen	University of North Carolina Chapel Hill, USA
Mads Neilsen	University of Copenhagen, Denmark
Wiro Niessen	Eramus MC Rotterdam, The Netherlands
Marc Niethammer	University of North Carolina Chapel Hill, USA
Xenophon Papademetris	Yale University, USA
Xavier Pennec	Inria, France
Joseine Pluim	Univeristy Medical Center Utrecht, The Netherlands
Killian Pohl	SRI International, USA
Francois Rousseau	CNRS, France
Daniel Rueckert	Imperial College London, UK
Mert Rory Sabuncu	Harvard Medical School, USA
Yoshinobu Sato	Osaka University, Japan
Julia Schnabel	University of Oxford, UK
Julia Schnabel	University of Oxford, UK
Dinggang Shen	University of North Carolina Chapel Hill, USA
Li Shen	Indiana University, USA
Kaleem Siddiqi	McGill University, Canada
Lawrence Staib	Yale University, USA
Danail Stoyanov	University College London, UK
Colin Studholme	University of Washington, USA
Gabor Szekely	ETH Zurich, Switzerland
Junichi Tokuda	Harvard Medical School, USA
Koen Van Leemput	Technical University of Denmark, Denmark
Rene Vidal	Johns Hopkins University, USA
Simon Warfield	Harvard Medical School, USA
William (Sandy) Wells	Harvard Medical School, USA
Carl-Fredrik Westin	Harvard Medical School, USA
Alistair Young	University of Auckland, New Zealand

Additional Reviewers

Abu Anas, Emran
Abugharbieh, Rafeef
Acar, Burak
Acosta-Tamayo, Oscar
Afacan, Onur
Afsari, Bijan
Aganj, Iman
Aja-Fernandez, Santiago
Akbari, Hamed
Akhondi-Asl, Alireza
Alavi, Abass
Alberola-Lopez, Carlos
Alexander, Daniel
Aljabar, Paul
Allan, Maximilian
Amini, Amir
Angelini, Elsa
Angelopoulou, Elli
Antani, Sameer
Anwander, Alfred
Arbel, Tal
Arbelaez, Pablo
Ardekani, Siamak
Arimura, Hidetaka
Arteta, Carlos
Ashburner, John
Ashraf, Ahmed
Aubert-Broche, Berengere
Audette, Michel
Auvray, Vincent
Avants, Brian
Avants, Brian
Awate, Suyash
Bach Cuadra, Meritxell
Bach Cuadra, Meritxell
Bagci, Ulas
Baka, Nora
Balicki, Marcin
Baloch, Sajjad
Barbu, Adrian
Barmpoutis, Angelos
Bartoli, Adrien
Bassett, Danielle
Batmanghelich, Kayhan
Bauer, Stefan
Baust, Maximilian
Bazin, Pierre-Louis
Bejar, Benjamin
Ben Ayed, Ismail
Bergeles, Christos
Berger, Marie-Odile
Bernardis, Elena
Betrouni, Nacim
Bharat, Shyam
Bhatia, Kanwal
Bhotika, Rahul
Bilgic, Berkin
Birkfellner, Wolfgang
Biros, George
Biros, George
Bismuth, Vincent
Boctor, Emad
Bogunovic, Hrvoje
Boisvert, Jonathan
Bossa, Matias Nicolas
Bouix, Sylvain
Boukerroui, Djamal
Bourgeat, Pierrick
Bovendeerd, Peter
Brady, Michael
Bron, Esther
Brost, Alexander
Buelow, Thomas
Buhmann, Joachim
Butakoff, Constantine
Caan, Matthan
Cabrera Lozoya, Roco
Cahill, Nathan
Cai, Weidong
Calder, Jeffrey
Camara, Oscar
Carass, Aaron
Cardenes, Ruben
Cardoso, Manuel Jorge
Carmichael, Owen
Caruyer, Emmanuel

Castellani, Umberto
Cathier, Pascal
Cattin, Philippe C.
Cepek, Jeremy
Ceresa, Mario
Cetingul, Hasan Ertan
Chakravarty, M. Mallar
Chang, Ping-Lin
Chaudhry, Rizwan
Chefd'hotel, Christophe
Chen, Elvis C. S.
Chen, Terrence
Chen, Thomas Kuiran
Chen, Ting
Cheng, Jian
Cheng, Jun
Cheriet, Farida
Chinzei, Kiyoyuki
Christensen, Gary
Chung, Albert C. S.
Chung, Moo
Cinquin, Philippe
Ciompi, Francesco
Ciuciu, Philippe
Clancy, Neil T.
Claridge, Ela
Clarkson, Matthew
Clarysse, Patrick
Claus, Piet
Clemmesen, Line
Cobzas, Dana
Collins, D. Louis
Colliot, Olivier
Commowick, Olivier
Cook, Philip
Cootes, Tim
Cordier, Nicolas
Corso, Jason
Counsell, Serena J.
Coupe, Pierrick
Cowan, Brett
Crane, Jason
Criminisi, Antonio
Crum, William
Cuingnet, Remi
Daga, Pankaj
Dalca, Adrian
Darkner, Sune
Das, Sandhitsu
Dauguet, Julien
Dawant, Benoit
De Craene, Mathieu
De Raedt, Sepp
Dehghan, Ehsan
Deligianni, Fani
Delingette, Herve
Demirci, Stefanie
Denney, Tom
Dequidt, Jeremie
Descoteaux, Maxime
Desjardins, Adrien
D'hooge, Jan
Di Battista, Andrew
Dijkstra, Jouke
DiMaio, Simon
Ding, Kai
Dojat, Michel
Donner, Rene
Drew, Mark
Du Bois d'Aische, Aloys
Duchateau, Nicolas
Duchesnay, Edouard
Duchesne, Simon
Duda, Jeffrey
Duits, Remco
Duriez, Christian
Dzyubachyk, Oleh
Eavani, Harini
Ebrahimi, Mehran
Edwards, Philip
Ehrhardt, Jan
Eklund, Anders
El-Baz, Ayman
Elson, Daniel
El-Zehiry, Noha
Erdt, Marius
Erus, Guray
Fahrig, Rebecca
Falco, Alexandre
Fallavollita, Pascal

Fang, Ruogu
Farag, Aly
Fedorov, Andriy
Fenster, Aaron
Feragen, Aasa
Figl, Michael
Fishbaugh, James
Fitzpatrick, J. Michael
Fletcher, P. Thomas
Florack, Luc
Foroughi, Pezhman
Fradkin, Maxim
Freiman, Moti
Freysinger, Wolfgang
Fripp, Jurgen
Fritscher, Karl
Fua, Pascal
Fuerst, Bernhard
Funka-Lea, Gareth
Funke, Jan
Gangeh, Mehrdad
Ganz, Melanie
Gao, Fei
Gao, Mingchen
Gao, Wei
Gao, Yaozong
Gao, Yi
Gaonkar, Bilwaj
Garcia-Lorenzo, Daniel
Garvin, Mona
Gaser, Christian
Ge, Tian
Georgescu, Bogdan
Geremia, Ezequiel
Ghanbari, Yasser
Gholipour, Ali
Ghosh, Aurobrata
Ghosh, Satrajit
Ghosh, Subham
Giannarou, Stamatia
Gibaud, Bernard
Gibson, Eli
Gilbert, Stephen
Gilles, Benjamin
Ginsburg, Shoshana
Giusti, Alessandro
Glocker, Ben
Goh, Alvina
Goksel, Orcun
Goldberger, Jacob
Goni Cortes, Joaquin
Gooya, Ali
Graham, Jim
Grbic, Sasa
Grisan, Enrico
Grova, Christophe
Guetter, Christoph
Guevara, Pamela
Gulsun, Mehmet Akif
Guo, Yanrong
Gur, Yaniv
Gutman, Boris
Hacihaliloglu, Ilker
Haidegger, Tamas
Haj-Hosseini, Neda
Hajnal, Joseph
Hamamci, Andac
Hammers, Alexander
Han, Lianghao
Haneishi, Hideaki
Hastreiter, Peter
Hatt, Chuck
Hauta-Kasari, Markku
Hawkes, David
Haynor, David
He, Huiguang
He, Tiancheng
Heckemann, Rolf
Heimann, Tobias
Heinrich, Mattias Paul
Helmstaedter, Moritz
Heng, Pheng Ann
Hennemuth, Anja
Hermosillo, Gerardo
Hibar, Derrek
Hipwell, John
Ho, Harvey
Holmes, David
Holmes, Jeff
Hong, Byung-Woo

Hontani, Hidekata
Hoogendoorn, Corne
Hu, Chenhui
Hu, Mingxing
Hu, Yipeng
Hu, Zhihong
Hua, Xue
Huang, Heng
Huang, Junzhou
Huang, Xiaojie
Huang, Xiaolei
Huisman, Henkjan
Ibrahim, El-Sayed
Iglesias, Juan Eugenio
Ingalhalikar, Madhura
Iordachita, Iulian
Jacobs, Colin
Jafari-Khouzani, Kourosh
Jagadeesan, Jayender
Jain, Saurabh
Jannin, Pierre
Janoos, Firdaus
Janowczyk, Andrew
Ji, Shuiwang
Jiang, Menglin
Jiang, Yifeng
Jolly, Marie-Pierre
Jomier, Julien
Jones, Geoffrey
Jordan, Petr
Joshi, Anand
Joshi, Gopal Datt
Joshi, Sarang
Joung, Sanghyun
Kabus, Sven
Kachelrie, Marc
Kadoury, Samuel
Kahl, Fredrik
Kainmueller, Dagmar
Kakadiaris, Ioannis
Kapoor, Ankur
Kapur, Tina
Karacali, Bilge
Karssemeijer, Nico
Keeve, Erwin
Kelm, Michael
Kerrien, Erwan
Khallaghi, Siavash
Khan, Ali
Kharazmi, Pegah
Kherif, Ferath
Khurd, Parmeshwar
Kim, Boklye
Kim, Minjeong
Kindlmann, Gordon
King, Andrew
Kiraly, Atilla
Kirisli, Hortense
Kitasaka, Takayuki
Klein, Stefan
Klein, Tassilo
Klinder, Tobias
Knutsson, Hans
Koizumi, Norihiro
Kowalewski, Timothy
Krause, Oswin
Krieger, Axel
Kunz, Manuela
Kwitt, Roland
Kwon, Dongjin
Lotjonen, Jyrki
Ladikos, Alexander
Laine, Andrew
Lam, Fan
Lamata, Pablo
Landman, Bennett
Lang, Pencilla
Langerak, Thomas
Langs, Georg
Lapeer, Rudy
Larsen, Anders Boesen Lindbo
Lartizien, Carole
Lasser, Tobias
Lasso, Andras
Lauze, Francois
Law, Max W.K.
Le, Yen
Lecoeur, Jeremy
Lee, Junghoon
Lee, Su-Lin

Lefevre, Julien
Lefkimmiatis, Stamatis
Le Folgoc, Loc
Lekadir, Karim
Lelieveldt, Boudewijn
Lendvay, Thomas
Lenglet, Christophe
Lepore, Natasha
Lesage, David
Li, Chunming
Li, Fuhai
Li, Gang
Li, Jiang
Li, Quanzheng
Li, Yang
Liao, Rui
Lillholm, Martin
Lin, Henry
Lindner, Claudia
Linguraru, Marius George
Linte, Cristian
Litjens, Geert
Liu, David
Liu, Jiamin
Liu, Jianfei
Liu, Sidong
Liu, Xiaoxiao
Liu, Yixun
Lo, Benny
Lombaert, Herve
Lorenz, Cristian
Lorenzi, Marco
Lu, Le
Lu, Shijian
Lu, Xiaoguang
Luan, K.
Lui, Lok Ming
Luo, Xiongbiao
Lyksborg, Mark
Muller, Henning
Machiraju, Raghu
Maddah, Mahnaz
Madooei, Ali
Mahapatra, Dwarikanath
Mahdavi, Seyedeh Sara
Maier-Hein, Lena
Mailhe, Boris
Majumdar, Angshul
Malandain, Gregoire
Malgouyres, Francois
Manduca, Armando
Manjon, Jose V.
Manniesing, Rashindra
Mansi, Tommaso
Marchal, Maud
Marchesseau, Stephanie
Margeta, Jan
Mari, Jean-Martial
Mariottini, Gian Luca
Marsland, Stephen
Marti, Robert
Martel, Anne
Martin-Fernandez, Marcos
Masamune, Ken
Masutani, Yoshitaka
Mateus, Diana
Mattes, Julian
McClelland, Jamie
McCulloch, Andrew
McIntosh, Chris
Mcleod, Kristin
Medrano-Gracia, Pau
Mendrik, Adrienne
Menze, Bjoern
Metz, Coert
Meyer, Chuck
Michailovich, Oleg
Miga, Michael
Mihalef, Viorel
Miller, James
Miller, Karol
Mirzaalian, Hengameh
Mirzaee, Hanieh
Modat, Marc
Moghari, Mehdi
Mohamed, Ashraf
Mohareri, Omid
Momayyez, Parya
Montillo, Albert
Moore, John

Mountney, Peter
Murphy, Keelin
Nabavi, Arya
Najman, Laurent
Nakamura, Ryoichi
Nakamura, Yoshihiko
Nasiriavanaki, Mohammadreza
Navab, Nassir
Neumuth, Thomas
Ng, Bernard
Nguyen, Hian
Nichols, Thomas
Nicolau, Stephane
Nie, Jingxin
Niederer, Steven
Nielsen, Poul
Noble, Alison
Noble, Jack
Noblet, Vincent
Nolte, Lutz
Nordsletten, David
Nouranian, Saman
Nugroho, Hermawan
Oda, Masahiro
O'Donnell, Lauren
O'Donnell, Thomas
Ofli, Ferda
Oliver, Arnau
Onofrey, John
Orihuela-Espina, Felipe
Otake, Yoshito
Ou, Yangming
Pace, Danielle
Padfield, Dirk
Padoy, Nicolas
Pallavaram, Srivatsan
Pant, Sanjay
Parisot, Sarah
Park, JinHyeong
Parthasarathy, Vijay
Pasternak, Ofer
Patriciu, Alexandru
Paul, Perrine
Paulsen, Rasmus
Pauly, Olivier
Pavlidis, Ioannis
Peitgen, Heinz-Otto
Penney, Graeme
Pernus, Franjo
Peterlik, Igor
Peters, Jochen
Peters, Terry M.
Petersen, Jens
Petersen, Jens
Petersen, Kersten
Petitjean, Caroline
Peyrat, Jean-Marc
Pham, Dzung
Piella, Gemma
Pitiot, Alain
Piuze, Emmanuel
Pizer, Stephen
Plenge, Esben
Poline, Jean-Baptiste
Poot, Dirk
Pop, Mihaela
Poulsen, Catherine
Prasad, Gautam
Prastawa, Marcel
Pratt, Philip
Prevost, Raphael
Price, Brian
Price, True
Prince, Jerry
Punithakumar, Kumaradevan
Qazi, Arish A.
Qi, Jinyi
Qian, Xiaoning
Qian, Zhen
Qiu, Wu
Radeva, Petia
Rafii-Tari, Hedyeh
Raj, Ashish
Rajagopalan, Vidya
Rajchl, Martin
Rajpoot, Nasir
Raju, Balasundar
Ramezani, Mahdi
Rapaka, Saikiran
Rathi, Yogesh

Ravichandran, Avinash
Reader, Andrew
Reiley, Carol
Reinertsen, Ingerid
Reisert, Marco
Reiter, Austin
Reyes, Mauricio
Rhode, Kawal
Richa, Rogrio
Riddell, Cyrill
Riklin Raviv, Tammy
Risholm, Petter
Risser, Laurent
Rit, Simon
Rivaz, Hassan
Riviere, Denis
Robert, Jean-luc
Roche, Alexis
Rohlfing, Torsten
Rohling, Robert
Rohr, Karl
Roth, Holger
Roysam, Badrinath
Russakoff, Daniel
Sorensen, Lauge
Saad, Ahmed
Sakuma, Ichiro
Salcudean, Tim
Salvado, Olivier
San Jose Estepar, Raul
Sanchez, Clarisa
Sands, Greg
Sarrut, David
Sarry, Laurent
Sarunic, Marinko
Savadjiev, Peter
Schaap, Michiel
Scheinost, Dustin
Scherrer, Benoit
Schmidt, Frank
Schmidt-Richberg, Alexander
Schneider, Caitlin
Schroeder, Peter
Schuh, Andreas
Schultz, Thomas
Schweikard, Achim
Seiler, Christof
Seitel, Alexander
Sermesant, Maxime
Seshamani, Sharmishtaa
Shahzad, Rahil
Shamir, Reuben R.
Sharma, Puneet
Shi, Feng
Shi, Kuangyu
Shi, Wenzhe
Shi, Yonggang
Shi, Yonghong
Shinohara, Russell T.
Shou, Guofa
Sijbers, Jan
Simon, Duchesne
Simpson, Amber
Simpson, Ivor
Singh, Nikhil
Singh, Vikas
Singh, Vivek
Sivaswamy, Jayanthi
Smeets, Dirk
Song, Gang
Song, Qi
Sotiras, Aristeidis
Sparks, Rachel
Speidel, Stefanie
Sporring, Jon
Spottiswoode, Bruce
Stuhmer, Jan
Stamm, Aymeric
Staring, Marius
Stetten, George
Stewart, Charles
Stewart, James
Styles, Iain
Styner, Martin
Su, Baiquan
Subramanian, Navneeth
Suinesiaputra, Avan
Suk, Heung-Il
Summers, Ronald
Sundar, Hari

Syeda-Mahmood, Tanveer
Szkulmowski, Maciej
Sznitman, Raphael
Tahmasebi, Amir
Taimouri,Vahid
Tamaki, Toru
Tan, Chaowei
Tanner, Christine
Taquet, Maxime
Tasdizen, Tolga
Taylor, Russell
Taylor, Zeike
Tek, Huseyin
Thirion, Bertrand
Thodberg, Hans Henrik
Thompson, Stephen
Tiwari, Pallavi
Tobon-Gomez, Catalina
Toews, Matthew
Tosun, Duygu
Totz, Johannes
Toussaint, Nicolas
Treeby, Bradley
Troccaz, Jocelyne
Tustison, Nicholas
Twining, Carole
Ukwatta, Eranga
Unal, Gozde
Uzunbas, Mustafa
Vaillant, Regis
Van Assen, Hans
Van de Ven, Wendy
Van Ginneken, Bram
Van Rikxoort, Eva
Van Walsum, Theo
Vannier, Michael
Varoquaux, Gael
Vegas-Sanchez-Ferrero, Gonzalo
Vemuri, Baba
Venkataraman, Archana
Vercauteren, Tom
Vialard, Francois-Xavier
Vignon, Francois
Vik, Torbjorrn
Villard, Pierre-Frederic
Visentini-Scarzanella, Marco
Viswanath, Satish
Vitaladevuni, Shiv
Vitanovski, Dime
Vogelstein, Joshua
Voigt, Ingmar
Von Berg, Jens
Voros, Sandrine
Vos, Pieter
Vosburgh, Kirby
Vrooman, Henri
Vrtovec, Tomaz
Wachinger, Christian
Waelkens, Paulo
Wan, Catherine
Wang, Chaohui
Wang, Hongzhi
Wang, Junchen
Wang, Li
Wang, Liansheng
Wang, Linwei
Wang, Qiu
Wang, Shijun
Wang, Song
Wang, Vicky
Wang, Yalin
Wang, Yinhai
Wang, Yu-Ping
Washio, Toshikatsu
Wassermann, Demian
Wee, Chong-Yaw
Weese, Jurgen
Wei, Liu
Wein, Wolfgang
Wels, Michael
Werner, Rene
Wesarg, Stefan
Whitaker, Ross
Whitmarsh, Tristan
Wiles, Andrew
Wilson, Nathan
Wittek, Adam
Wolz, Robin
Wong, Ken C.L.
Wong, Stephen

Wright, Graham
Wu, Guorong
Wu, John Jue
Wu, Wen
Wu, Xiaodong
Xie, Yuchen
Xing, Fuyong
Xu, Xiao Yun
Xu, Yanwu
Xu, Ziyue
Xue, Zhong
Yan, Pingkun
Yan, Zhennan
Yang, Liangjing
Yang, Lin
Yaniv, Ziv
Yao, Jianhua
Yap, Pew-Thian
Yaqub, Mohammad
Ye, Dong Hye
Yeo, B.T. Thomas
Yin, Youbing
Yin, Zhaozheng
Yokota, Futoshi
Yoshida, Hiro
Yushkevich, Paul
Zeng, Haishan
Zeng, Wei
Zhan, Liang
Zhan, Yiqiang
Zhang, Daoqiang
Zhang, Hui
Zhang, Pei
Zhang, shaoting
Zhang, Zhijun
Zhao, Qian
Zheng, Guoyan
Zheng, Yefeng
Zheng, Yuanjie
Zhou, Jian
Zhou, Jinghao
Zhou, Kevin
Zhou, Luping
Zhou, X. Sean
Zhou, Yan
Zhu, Dajiang
Zhu, Hongtu
Zhu, Yuemin
Zhuang, Xiahai
Zijdenbos, Alex
Zikic, Darko
Zollei, Lilla
Zuluaga, Maria A.
Zwiggelaar, Reyer

MICCAI Society Board of Directors

Alison Noble (President)	University of Oxford, Oxford, UK
Wiro Niessen (Exec. Director)	Erasmus MC - University Medical Centre, Rotterdam, The Netherlands
Sebastien Ourselin (Treasurer)	University College, London, UK
Gabor Fichtinger (Secretary)	Queen's University, Kingston, ON, Canada
Stephen Aylward	Kitware, Inc., NY, USA
Alejandro Frangi	University of Sheffield, Sheffield, UK
Polina Golland	MIT, Cambridge, MA, USA
Pierre Jannin	INSERM, Rennes, France
Kensaku Mori	Nagoya University, Japan
Nassir Navab	TUM, Munich, Germany
Xavier Pennec	INRIA, Sophia Antipolis, France
Josien Pluim	University Medical Center Utrecht, The Netherlands
Daniel Rueckert	Imperial College, London, UK
Dinggang Shen	UNC, Chapel Hill, NC, USA

MICCAI Society Consultants to the Board

Alan Colchester	University of Kent, UK
Terry Peters	University of Western Ontario, London, Canada
Richard Robb	Mayo Clinic College of Medicine, USA

MICCAI Society Staff

Society Secretariat	Janette Wallace, Canada
Recording Secretary	Jackie Williams, Canada

Table of Contents – Part III

Shape and Population Analysis

Brain II

Diffusion MRI

Machine Learning II

Generalized Multiresolution Hierarchical Shape Models via Automatic Landmark Clusterization

Juan J. Cerrolaza[1], Arantxa Villanueva[2], Mauricio Reyes[3], Rafael Cabeza[2], Miguel Angel González Ballester[4], and Marius George Linguraru[1,5]

[1] Sheikh Zayed Institute for Pediatric Surgical Innovation, Children´s National Health System, Washington DC, USA
{JCerrola,MLingura}@cnmc.org
[2] Public University of Navarra, Pamplona, Spain
[3] Surgical Technology and Biomechanics, University of Bern, Bern, Switzerland
[4] ICREA, Barcelona, Spain – Universitat Pompeu Fabra, Barcelona, Spain
[5] School of Medicine and Health Sciences, George Washington Univ. Washington, DC, USA

Abstract. Point Distribution Models (PDM) are some of the most popular shape description techniques in medical imaging. However, to create an accurate shape model it is essential to have a representative sample of the underlying population, which is often challenging. This problem is particularly relevant as the dimensionality of the modeled structures increases, and becomes critical when dealing with complex 3D shapes. In this paper, we introduce a new generalized multiresolution hierarchical PDM (GMRH-PDM) able to efficiently address the high-dimension-low-sample-size challenge when modeling complex structures. Unlike previous approaches, our new and general framework extends hierarchical modeling to any type of structure (multi- and single-object shapes) allowing to describe efficiently the shape variability at different levels of resolution. Importantly, the configuration of the algorithm is automatized thanks to the new agglomerative landmark clustering method presented here. Our new and automatic GMRH-PDM framework performed significantly better than classical approaches, and as well as the state-of-the-art with the best manual configuration. Evaluations have been studied for two different cases, the right kidney, and a multi-object case composed of eight subcortical structures.

Keywords: Shape models, multiresolution, hierarchical models, PDM.

1 Introduction

Since their inception in the early 1990s, active shape models (ASM) [1] have proven effective for addressing a number of problems where the target structures are consistent in shape but poorly defined by image features, as is often the case in medical images. The success of point distribution models (PDM)-based matching approaches depends on an accurate description of the shape class, the expected shape instances, and their variations. While a limited number of examples may be sufficient when working with relatively simple objects, an adequately large training set is not always available as the dimensionality and complexity of the structures increase, as is usually

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 1–8, 2014.

the case when working with 3D multi-object structures. This issue is known as the high-dimension-low-sample-size (HDLSS) problem. Trying to overcome this question, some authors have proposed interesting versions of the classical PDM, exploiting the possibilities of incorporating multiresolution-based hierarchical analysis to shape modeling. Davatzikos et al. [2] proposed the hierarchical decomposition of the shape into small pieces of information via the wavelet transform. However, whereas the independent modeling of these bands allows reducing the dimensionality of the problem, and thus the HDLSS effect, it also reduces the robustness of the model as shown in [3]. An interesting attempt to describe the interrelationships between objects at different scales statistically is the multiscale framework proposed by Lu et al. [4], using m-reps as the geometric representation of shapes. In spite of the valuable multiscale properties of m-reps, they are less intuitive than the landmarks-based representation used in PDMs, which is probably one of the simplest and most generic methods used to represent shapes. Yokota et al. [13] proposed an interesting hierarchical statistical model of the femur and pelvis, imposing additional connectivity constraints to control the matching between different subparts. In the recent work of Cerrolaza et al. [3,5], a new multiresolution hierarchical variant of PDM (MRH-PDM) was introduced, able to efficiently characterize the different inter-object relationships, as well as the particular locality of each element separately. Even though the potential of this new method in terms of accuracy and robustness improvement was successfully verified, there are two main drawbacks that limit its practical application. First, the absence of an automatic grouping approach can hinder its use when working with complex data with a large number of objects. On the other hand, the hierarchical decomposition is limited to multi-object structures, since no intra-object analysis is considered within the original framework.

In this paper, we propose a new Generalized Multiresolutoin Hierarchical PDM (GMRH-PDM) that addresses these two important issues, automatic grouping and intra-object analysis. The new notation introduced in Section 2 extends the hierarchical modeling of PDM even to single-object structures, which leads to a more versatile and generalizable framework. Finally, the configuration of the algorithm (i.e., the definition of clusters at each resolution) is automatized thanks to the new agglomerative landmark clustering approach described in Section 3. The performance of the new GMRH-PDM method is studied for two different cases, the right kidney, and a multi-object case composed of eight subcortical structures.

2 Generalized Multiresolution Hierarchical PDM

In this section we present a new generalization of the original MRH-PDM formulation described by Cerrolaza et al. [3]. In their work the capability to model variability in subparts of a single object was limited, as they considered the single objects as the simplest structure to model at the finest resolution levels. Relaxing this condition, we go one step further in the development of hierarchical PDMs, introducing a more general framework where any possible grouping of landmarks is considered.

Let $\boldsymbol{x}$ be the vector form of a 3D shape defined by $K \in \mathbb{N}$ landmarks. In the general case of a multi-object shape composed of M ($M \in \mathbb{N}$) single-object structures, $\boldsymbol{x_j}$ ($1 \leq j \leq M$), $\boldsymbol{x}$ is defined by the concatenation of the 3 coordinates of the $K_j \in \mathbb{N}$ landmarks $\left(K = \sum K_j\right)$ that define each object, i.e. $\boldsymbol{x} = (\boldsymbol{x}_1; \ldots; \boldsymbol{x}_M)^{\boldsymbol{T}}$. Using the multi-object generalization [5] of the matrix notation initially proposed by Lounsbery et al. [6], the multiresolution analysis of $\boldsymbol{x}$ can be formulated as:

$$\boldsymbol{x}^r = \boldsymbol{A}^r \boldsymbol{x}^{r-1} \tag{1}$$

$$\boldsymbol{z}^r = \boldsymbol{B}^r \boldsymbol{x}^{r-1} \tag{2}$$

where $r \in \mathbb{N}$ indicates the level of resolution (in particular $r = 0$ defines the finest level of resolution, and thus, $\boldsymbol{x}^0 = \boldsymbol{x}$), and $\boldsymbol{A}^r$ and $\boldsymbol{B}^r$ represent the analysis filters. Equation (1) implements the filtering and downsampling of $\boldsymbol{x}^{r-1}$, providing a lower resolution version of it (i.e., $K^{r-1} > K^r$, where $K^r \in \mathbb{N}$ represents the number of landmarks at the resolution level r), while $\boldsymbol{z}^r$ (2) captures the lost detail between $\boldsymbol{x}^r$ and $\boldsymbol{x}^{r-1}$. An optimal selection of these analysis filters guarantees that no information is lost during the process, being possible to reverse the analysis process with synthesis equation: $\boldsymbol{x^{r-1}} = \boldsymbol{F}^r \boldsymbol{x}^r + \boldsymbol{G}^r \boldsymbol{z}^r$. Lounsbery et al. [6] provide a general multiresolution framework to compute the analysis and synthesis filters ($\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{F}$, and $\boldsymbol{G}$) for meshes with subdivision connectivity and arbitrary topology. In this work, we define the multiresolution domain using the octahedron as the reference mesh [8], with a 4-to-1 splitting step, and a lifter butterfly scheme for triangular meshes [7].

With this method described above, it is possible to decompose any multi object structure into different levels of resolution. Whereas MRH-PDM established a specific division of the M objects into $M^r \in \mathbb{N}$ disjoint subsets at each level or resolution (i.e., only complete objects can be part of a subset), here we propose a more general definition of the disjoint subsets allowing any type of groping between the total set of landmarks. Thus, at each level of resolution r we define a particular division of the K^r landmarks into M^r separate clusters, $\left(\boldsymbol{S}_1^r, \ldots, \boldsymbol{S}_{M_r}^r\right)$, where $\boldsymbol{S}_s^r$ ($s = 1, \ldots, M^r$) is formed by the indices of the landmarks contained in this subset, and therefore, $\bigcap_{s=1}^{M_r} \boldsymbol{S}_s^r = \emptyset$ and $\bigcup_{s=1}^{M_r} \boldsymbol{S}_s^r = (1, \ldots, M)$. In addition, we impose the following condition. Suppose $\boldsymbol{S}_{s^{r-1}}^{r-1}(i)$ represents the i-th element of the s^{r-1}-th subset defined at the r-1-th resolution level, and $\widehat{\boldsymbol{S}}_{s^r}^r$ be the propagation of $\boldsymbol{S}_{s^r}^r$ to r-1, then

$$\boldsymbol{S}_{\boldsymbol{s^{r-1}}}^{r-1}(i) \subseteq \widehat{\boldsymbol{S}}_{s^r}^r \implies \forall \boldsymbol{S}_{\boldsymbol{s^{r-1}}}^{r-1} \subseteq \widehat{\boldsymbol{S}}_{s^r}^r \tag{3}$$

That is, two sets of landmarks that have been grouped separately at a specific level of resolution, should not be jointly modeled at finer levels; or equivalently, the clusters created in the resolution r-th derived from the fragmentation of clusters in resolution $r + 1$-th (see Fig. 1). Despite the intuitive meaning of (3), there is a challenge yet to be resolved: the propagation of the clusters between two consecutive resolutions. Let $\boldsymbol{l}^r$ be a $(3K^r \times 1)$ vector (i.e., the same size as $\boldsymbol{x}^r$) containing the labels of the subset to which each landmark of $\boldsymbol{x}^r$ belongs; i.e., if $\boldsymbol{x}^r(i) \epsilon \boldsymbol{S}_s^r$ then $\boldsymbol{l}^r(i) = s$. With this notation, we can estimate $\hat{\boldsymbol{l}}^{r-1}$, the propagation of the subdivision defined by $\boldsymbol{l}^r$ to the landmarks of the following resolution, r-1, by means of the synthesis matrix, $\boldsymbol{F}^r$.

3 Automatic Landmark Clustering Using Vector Fields

In this section we introduce a new landmark clustering approach that allows to define automatically the division of the landmarks into separate clusters at each resolution. The clustering process was initially inspired by the work presented by Roy et al. [9], which was originally conducted for vector field segmentation of moving objects in 2D videos, and extended to 3D objects by Reyes et al. [10] to study the anatomical variability of single organs via principal factor analysis. Here we propose a more general approach based on the agglomerative hierarchical clustering method presented by Ward [11], where the criterion for choosing the pair of clusters to merge at each step is based on the minimum value of the tailored objective function:

$$J(\Omega) = \alpha_1 \int_{\Omega} \left(\frac{|\boldsymbol{V}_{\Omega} \times \boldsymbol{V}_i|}{|\boldsymbol{V}_i|}\right)^2 \frac{L_{max}}{|\boldsymbol{V}_i|} di + \alpha_2 \left(1 - \frac{\int_{\Omega} di}{\int_{\boldsymbol{S}} di}\right) + \alpha_3 H(\Omega) \tag{4}$$

where α_1, α_2 and α_3 are real values such that $\sum \alpha_i = 1$. $\Omega \subseteq \boldsymbol{S}$ represents a region or subdomain within the set of landmarks $\boldsymbol{S}$ we want to divide into an optimal set of clusters. The first component of (4) takes into account the colinearity between deformation vectors within the domain Ω and the predominant vector direction $\boldsymbol{V}_{\Omega}$ in Ω. Here, we define the deformation vector of landmark $l_i \in \boldsymbol{S}$, $\boldsymbol{V}_i$ as the sum of the eigenvectors obtained via PDM over $\boldsymbol{S}$, and weighted by their corresponding eigenvalues. Then $L_{max} = max_{\boldsymbol{S}}\{\|\boldsymbol{V}_i\|\}$, and $\boldsymbol{V}_{\Omega}$ is defined as the highest eigenvalue of the matrix $M(\Omega) = \int_{\Omega} \boldsymbol{V}_i \, \boldsymbol{V}_i{}^t di$. The second term in (4) acts as a maximal area constraint. The aim of the third term, $H(\Omega)$, defined as the Hausdorff distance between the objects that compose Ω normalized by the maximum distance among objects in $\boldsymbol{S}$, is to promote the grouping of objects that are spatially close. When minimizing equation (4), it is desirable that the colinearity between deformation vectors be the dominant term in the generation of clusters, while the second and third term act as additional constraint to guarantee the consistency of the final results, i.e. $\alpha_1 \gg (\alpha_2, \alpha_3)$.

From the family of partitions provided by Ward's [11] algorithm, we define the optimal landmark division based on a tailored version of the Silhouette coefficient defined below. Suppose landmark l_i is assigned to cluster Ω_i. Then, it is possible to define how well l_i is assigned to its cluster as $a_i = J(\Omega_i) - J(\Omega_{i \setminus l_i})$, where $\Omega_{i \setminus l_i}$ represents the cluster Ω_i after removing l_i. Thus, large a_i represents a high dissimilarity between l_i and Ω_i. In the same way, we define the dissimilarity of l_i to any other cluster $\Omega_j (j \neq i)$ to which l_i is not member as $b_i = min\{b_{i,j}\}$, where $b_{i,j} = J(\Omega_{j+l_i}) - J(\Omega_j)$, and Ω_{j+l_i} represents the union of l_i and Ω_j. Constraining the value of a_i and b_i to the range $[0,1]$ by means of the logistic function, $LF(\cdot)$, we define the Silhouette coefficient for landmark l_i, s_i, as

$$s_i = \frac{LF(b_i) - LF(a_i)}{max\{LF(b_i),\ LF(a_i)\}} \tag{5}$$

Since a value of s_i close to 1 means that l_i is appropriately clustered in Ω_i, the optimal clustering of $\boldsymbol{S}$ will be the one that maximizes the average s_i.

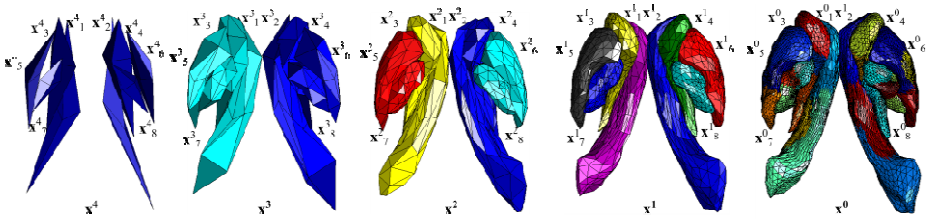


Fig. 1. Hierarchical configuration provided by the GMRH-PDM algorithm for the multi-object structure composed by eight subcortical structures. At each level of resolution the set of landmarks depicted with the same color are modeled jointly via PDM. At resolution $\boldsymbol{x}^1$ the lateral ventricles are in magenta ($\boldsymbol{x}_1^1$) and navy ($\boldsymbol{x}_2^1$), the caudate nuclei in yellow ($\boldsymbol{x}_3^1$) and green ($\boldsymbol{x}_4^1$), the putamen in black ($\boldsymbol{x}_5^1$) and red ($\boldsymbol{x}_6^1$), and the globus pallidi in blue ($\boldsymbol{x}_7^1$) and cyan ($\boldsymbol{x}_8^1$).

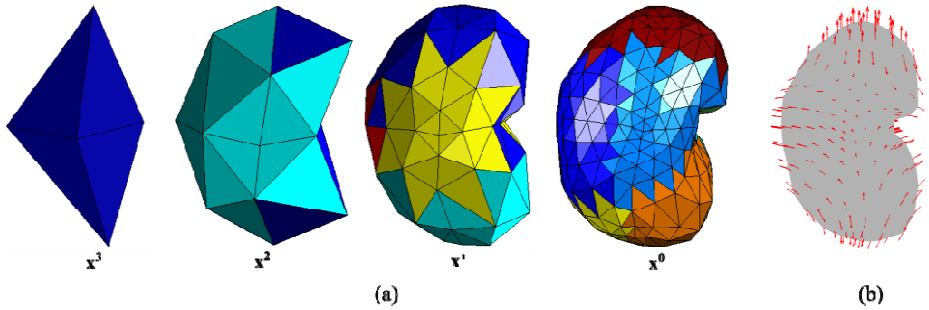


Fig. 2. (a) Hierarchical configuration obtained for the kidney model. At each level of resolution the area depicted with the same color is modeled jointly via PDM. (b) Deformation field defined by the first mode of variation when modeling all the landmarks jointly. The clustering configurations proposed by GMRH-PDM (especially in resolutions $\boldsymbol{x}^1$ and $\boldsymbol{x}^0$) are able to identify areas affected by a similar deformation field.

4 Shape Modeling via GMRH-PDM

Let $\boldsymbol{x}$ be the vector form of any 3D structure (i.e., multi-object or single-object), whose multiresolution decomposition $\{\boldsymbol{x} = \boldsymbol{x}^0, \boldsymbol{x}^1, \dots, \boldsymbol{x}^R, \boldsymbol{z}^1, \dots, \boldsymbol{z}^R\}$, can be obtained using (1) and (2). Imposing the initial condition that $M^R = 1$ (i.e., a global statistical shape model of the whole set is built at the coarsest resolution in order to guarantee the coherent disposition of the elements), a new landmark subdivision scheme is calculated at resolution r-1 for each of the M^rsubsets ($\boldsymbol{S}_s^r, s = 1, \dots, M^r$) obtained at r. Every new subdivision is obtained automatically using the landmark clustering approach introduced in Section 3. Finally, an efficient statistical model of the shape is created building a different PDM for every $\boldsymbol{S}_s^{r-1}$ in which the structure has been divided, allowing to characterize different characteristics of the structure at each scale. Suppose now we want to use the new GMRH-PDM we just created to describe a new case, $\boldsymbol{y}$, i.e., finding the best approximation of $\boldsymbol{y}$ in the subspace of allowed shapes described by the statistical model. Starting from the finest resolution, $\boldsymbol{y}^0$ is divided into the M^0 subsets previously defined, each of which is corrected by

the corresponding PDM. This process is repeated at each resolution until r = R. In the transition of each resolution, the high frequency component of the new constrained shape, $\hat{\boldsymbol{z}}^1$, will be used to recover the original resolution at the end of the process using the synthesis equation presented in Section 2.

5 Results and Discussion

To evaluate the performance of the new automatic GHMR-PDM approach we use two different datasets. First we use a set of 18 T1-weighted brain MRI volumes obtained from the Internet Brain Segmentation Repository (IBSR) [12] (pixel resolution $0.94 \times 0.94 \times 1.5$ mm; volumes: $256 \times 256 \times 128$ voxels). In particular, we work with a multi-object structure composed of eight subcortical structures ($\boldsymbol{x}_1, \ldots, \boldsymbol{x}_8$), corresponding to the left and right lateral ventricles, left and right caudate nuclei, left and right putamens, and left and right globus pallidi, respectively (Fig. 1). The performance of the new GMRH-PDM is also tested over a single-object database. We use a proprietary dataset of right kidneys from 18 CT abdominal studies (pixel resolution: $0.58 \times 0.58 \times 1.00$ mm; $512 \times 512 \times 360$ voxels). Following the general guidelines described in Section 3, the three configuration parameters of GMRH-PDM, α_1, α_2 and α_3, are set to 0.8, 0.1 and 0.1, respectively. Experimentally, we observed great similarity between the clusters obtained when α_1= [0.7 – 0.9] (using $\alpha_2 = \alpha_3 = (1 - \alpha_1)/2$). For $\alpha_1 < 0.7$, the landmarks grouped into a single large cluster, being the second and third term of (4) which control the clusterization process. For $\alpha_1 > 0.9$ landmarks are over-clustered due to the under-penalization of partitions.

The resulting automatic configurations for the multi-object and single-object case are shown in Figs. 1 and 2 respectively. The behavior of the new modeling approach is compared with two alternative methods for the multi-object case: the classical PDM [1], and the previous multiresolution hierarchical approach, MRH-PDM, proposed in [3, 5]. In particular, we chose the configuration that exhibited best results from all the hierarchical configurations manually defined in [5]. Due to the inability of MRH-PDM to deal with single-object structures, only PDM is considered in the comparison for the second/kidney data under study. The accuracy of the different methods to model new instances of the underlying population is evaluated in terms of the average landmark-to-landmark distance (L2L), and the Dice coefficient (DC), using leave-one-out cross-validation. Table 1 shows the results obtained for the multi-object case. Compared with the classical PDM (avg. L2L error: 1.20 ± 0.49 vox.; avg. DC: 0.87 ± 0.06), both multiresolution hierarchical approaches provide substantial improvements in accuracy for all the subcortical structures. With the exception of globus pallidi, all improvements over PDM are statistically significant according to the Wilcoxon signed rank test (p-value < 0.05 for all). Although the new GMRH-PDM performed similarly to the previous hierarchical version, MRH-PDM, in terms of accuracy (avg. L2L error: 0.94 ± 0.19 vs. 0.95 ± 0.39 and avg. DC: 0.90 ± 0.04 vs. 0.90 ± 0.05, respectively), it provides a significant advantage over the latter. The GMRH-PDM framework introduced in this paper is fully automatic, while the original MRH-PDM requires the hierarchical configuration to be manually defined by the

user. As the number of possible configurations can be considerably high when working with large number of objects, it is a nontrivial challenge to find an optimal one by simple manual supervised selection. Thanks to the landmark clustering approach presented in Section 3, GMRH-PDM is able to automatically provide an optimal hierarchical decomposition of the structure, while performing as well as the best manual configuration of MRH-PDM. But GMRH-PDM has an additional major advantage over MRH-PDM as it allows single-object and intra-object analysis.

Table 1. Accuracy Evaluation. Landmark-to-landmark (L2L) distance and Dice's coefficient (DC) (average / standard deviation) for the three studied methods (PDM, MRH-PDM, and GMRH-PDM) over eight subcortical structures ($\boldsymbol{x}_1, \dots, \boldsymbol{x}_8$) (see Fig. 1). (•) marks significant improvements over classic PDM.

L2L (vox.)	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$	$\boldsymbol{x}_7$	$\boldsymbol{x}_8$	Avg.
PDM	1.63/0.41	1.60/0.56	1.30/0.69	1.21/0.39	1.13/0.30	0.97/0.19	0.86/0.24	0.93/0.25	1.20/0.49
MRH-PDM	• 1.28/0.50	• 1.24/0.49	• 0.98/0.48	• 0.86/0.37	• 0.88/0.18	• 0.77/0.20	0.79/0.21	0.82/0.19	• 0.95/0.39
GMRH-PDM	• 1.26/0.45	• 1.23/0.50	• 0.97/0.47	• 0.85/0.36	• 0.88/0.17	• 0.77/0.19	0.78/0.20	0.82/0.20	• 0.94/0.19

DC	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$	$\boldsymbol{x}_7$	$\boldsymbol{x}_8$	Avg.
PDM	0.80/0.05	0.80/0.07	0.86/0.04	0.86/0.04	0.89/0.02	0.91/0.02	0.91/0.02	0.90/0.03	0.87/0.06
MRH-PDM	• 0.84/0.05	• 0.85/0.06	• 0.88/0.04	• 0.90/0.04	• 0.92/0.02	• 0.93/0.02	0.91/0.02	0.91/0.02	• 0.90/0.05
GMRH-PDM	• 0.85/0.04	• 0.87/0.06	• 0.89/0.03	• 0.90/0.03	• 0.93/0.03	• 0.94/0.02	0.91/0.02	0.92/0.03	• 0.90/0.04

The superiority of GMRH-PDM over PDM to model subparts in single-object structures was also proven in the kidney database. In this case, the average L2L errors were 0.35 ± 0.19 vs. 0.47 ± 0.2, and the average DCs were 0.99 ± 0.05 vs. 0.97 ± 0.05 for GMRH-PDM and PDM, respectively (p-value = 0.03 and 0.02 respectively).

The computational complexity of the new landmark clusterization is $O(n^2)$, taking ~ 100 min. to process the most complex multi-object case with 8208 landmarks (code written in Matlab®). However, this is not a determining factor for the practical application of the method, since the clusterization can be performed off-line.

6 Conclusions

In this paper, we present a new Generalized Multiresolutoin Hierarchical PDM (GMRH-PDM) to address the high-dimension-low-sample-size challenge of great relevance when modeling complex structures with the classical PDM. The general framework introduced here creates different statistical models that allow to describe efficiently the variability of the shape at different levels of resolution. The new GMRH-PDM also tackles the two main drawbacks observed in previous hierarchical approaches: the difficulty of manually defining the hierarchical configuration that provides optimal performance, and the impossibility of dealing with single-object structures by considering entire objects as the minimum modeling unit. The general notation used in GMRH-PDM extends the hierarchical modeling of PDM to any set of

landmarks, leading to a more versatile framework able to deal with all types of structures, even single-object shapes. Finally, the hierarchical configuration of the algorithm is automatically defined by means of a new agglomerative landmark clustering approach, whose optimization is controlled by a tailored definition of the Silhouette coefficient. The algorithm is compared with two different alternatives, PDM and the MRH-PDM. The results show how the new automatic GMRH-PDM significantly outperform the classical PDM in terms of accuracy, while providing similar results to the best manual configuration of MRH-PDM. GMRH-PDM allows the automatic hierarchical modeling of structures, from the multi-object level to the inter- and intra-object resolution, which can be of great interest in the context of full body computational anatomy. In the near future, we plan to continue exploring this capability to study population variability and the temporal anatomical variability of organs.

Acknowledgment. This project was supported by a philanthropic gift from the Government of Abu Dhabi to Children's National Health System, and by the European Union FP7 project HEAR-EU (grant agreement 304857).

References

1. Cootes, T.F., et al.: Active Shape ModelsTheir Training and Application. Comput. Vis. Image Underst. 61(1), 38–59 (1995)
2. Davatzikos, C., et al.: Hierarchical active shape models, using the wavelet transform. IEEE Trans. on Med. Imag. 22(3), 414–423 (2003)
3. Cerrolaza, J.J., et al.: Hierarchical Statistical Shape Models of Multiobject Anatomical Structures: Application to Brain MRI. IEEE Trans. Med. Imaging 31(3), 71–724 (2012)
4. Lu, C., et al.: Statistical multi-object shape models. Int. J. Comput. Vis. 75, 387–404 (2007)
5. Cerrolaza, J.J., Herrezuelo, N.C., Villanueva, A., Cabeza, R., González Ballester, M.A., Linguraru, M.G.: Multiresolution Hierarchical Shape Models in 3D Subcortical Brain Structures. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part II. LNCS, vol. 8150, pp. 641–648. Springer, Heidelberg (2013)
6. Lounsbery, M., et al.: Multiresolution Analysis for Surfaces of Arbitrary Topological Type. ACM Trans. Graph. 16(1), 34–73 (1997)
7. Dyn, N., et al.: A Butterfly Subdivision Scheme for Surface Interpolation with Tension Control. ACM Trans. Graph. 9(2), 160–169 (1990)
8. Praun, E., Hoppe, H.: Spherical Parametrization and Remeshing. ACM Trans. Graph. (SIGGRAPH) 22(3), 340–349 (2003)
9. Roy, T., et al.: Segmentation of a vector field: dominant parameter and shape optimization. Journal of Mathematical Imaging and Vision 24(2), 259–276 (2006)
10. Reyes, M., et al.: Anatomical variability of organs via principal factor analysis from the construction of an abdominal probabilistic atlas. IEEE Symp. Bio. Imag., 682–685 (2009)
11. Ward, J.H.: Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 58, 236–244 (1963)
12. IBSR. The Internet Brain Segmentation Repository (IBSR), `http://www.cma.mgh.harvard.edu/ibsr/`
13. Yokota, F., Okada, T., Takao, M., Sugano, N., Tada, Y., Sato, Y.: Automated segmentation of the femur and pelvis from 3D CT data of diseased hip using hierarchical statistical shape model of joint structure. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C., et al. (eds.) MICCAI 2009, Part II. LNCS, vol. 5762, pp. 811–818. Springer, Heidelberg (2009)

Hierarchical Bayesian Modeling, Estimation, and Sampling for Multigroup Shape Analysis

Yen-Yun Yu[1], P. Thomas Fletcher[1], and Suyash P. Awate[1,2]

[1] Scientific Computing and Imaging (SCI) Institute, School of Computing, University of Utah
[2] Computer Science and Engineering Department, Indian Institute of Technology (IIT), Bombay

Abstract. This paper proposes a novel method for the analysis of anatomical shapes present in biomedical image data. Motivated by the natural organization of population data into multiple groups, this paper presents a novel *hierarchical generative* statistical model on shapes. The proposed method represents shapes using pointsets and defines a joint distribution on the population's (i) shape variables and (ii) object-boundary data. The proposed method solves for optimal (i) point locations, (ii) correspondences, and (iii) model-parameter values as a *single* optimization problem. The optimization uses expectation maximization relying on a novel Markov-chain Monte-Carlo algorithm for *sampling* in Kendall shape space. Results on clinical brain images demonstrate advantages over the state of the art.

Keywords: Shape analysis, hierarchical Bayes, sampling in shape space.

1 Introduction and Related Work

Shape analysis [6,9] entails the inference of shape models from population data and associated statistical analyses, e.g., hypothesis testing for comparing groups. The natural organization of biomedical data into groups, and possibly subgroups, calls for a *hierarchical* modeling strategy. Previous works on hierarchical shape modeling typically concern (i) multi-resolution models, e.g., a face model at fine-to-coarse resolutions, or (ii) multi-part models, e.g., a car decomposed into body, tires, and trunk. In contrast, the proposed framework deals with population data comprising multiple groups, e.g., the Alzheimer's disease (AD) population comprising people with (i) dementia due to AD, (ii) mild cognitive impairment due to AD, and (iii) preclinical AD.

Figure 1 outlines the proposed *generative* model, where (i) top-level variables capture the shape properties across the population (e.g., all individuals with and without medical conditions), (ii) variables at a level below capture the shape distribution in different groups within the population (e.g., clinical cohorts based on gender or type of disease within a spectrum disorder), and (iii) variables at the next lower level capture individual shapes, which finally relate to (iv) individual image data at the lowest level. Moreover, the top-level population variables provide a common reference frame for the group shape models, which is necessary to enable comparison between the groups.

This paper makes several contributions. **(I)** It proposes a novel hierarchical generative model for population shape data. It represents a shape as an equivalence class of pointsets modulo translation, rotation, and isotropic scaling [6]. This model tightly couples each individual's shape (unknown) to the observed image data by designing their

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 9–16, 2014.

joint probability density function (PDF) using current distance or kernel distance [8,17]. The current distance makes the logarithm of the joint PDF a nonlinear function of the point locations. Subsequently, the proposed method solves a *single unified model-fitting optimization problem to estimate optimal point locations, correspondences, and parameter values*. **(II)** The proposed model fitting relies on expectation maximization (EM), treating the individual-shape and group-shape variables as hidden random variables, thereby integrating them out while estimating parameters (e.g., the population shape mean and covariance). In this way, the proposed EM algorithm improves over typical methods that use mode approximation for shape variables. **(III)** The EM algorithm entails evaluating an expectation over the posterior PDF of the shape variables. For instance, the posterior PDF for individual-shape variables involves the (i) likelihood PDF designed using the current distance and (ii) prior PDF conditioned on the group shape model. To compute the expectation, the proposed EM algorithm relies on a novel adaptation of Hamiltonian Monte Carlo (HMC) [5] *sampling in Kendall shape space*. **(IV)** The results show that the hierarchical model leads to more compact model fits and improved detection of subtle shape variations between groups.

Early approaches [2,6] to statistical shape modeling rely on manually defined homologous landmarks. Later approaches optimize point positions or correspondences using statistical compactness criteria such as the (i) logarithm of the determinant of the model covariance matrix [10], (ii) minimum description length [4,16], or (iii) minimum entropy [1]. However, these approaches (i) do *not* incorporate a generative statistical model, (ii) introduce adhoc terms in the objective function to obtain correspondences, and (iii) do *not* estimate shape-model parameters within the aforementioned optimization. Some generative models for shape analysis do exist [3,7,12,14], but these models rely on a pre-determined template shape with manually placed landmarks.

2 Hierarchical Bayesian Shape Model

We first describe the proposed hierarchical model for multigroup shape data (Figure 1).

Data: Consider a group of I vector random variables $X := \{X_i\}_{i=1}^{I}$, where X_i is a vector random variable denoting a given set of points on the boundary of an anatomical structure in the i-th individual's image data. That is, $X_i := \{X_i(n)\}_{n=1}^{N_i}$ where $X_i(n) \in \mathbb{R}^D$ is the D-dimensional spatial coordinate of the n-th point in the pointset. Such points can be obtained from a given segmentation or delineation of the anatomical structure. In this paper, $D = 3$. In any individual's image data, the number of boundary points N_i can be arbitrary. Similarly, consider other groups of data, e.g., data $Y := \{Y_j\}_{j=1}^{J}$ derived from a group of J individuals, data $\{Z_k\}_{k=1}^{K}$, etc.

Individual Shape Variables: For the first group (corresponding to data X), consider a group of I *hidden* random variables $U := \{U_i\}_{i=1}^{I}$, where U_i is a vector random variable representing the shape of the anatomical structure of the i-th individual. That is, $U_i := \{U_i(t)\}_{t=1}^{T}$ where $U_i(t) \in \mathbb{R}^D$ is the D-dimensional coordinate of the t-th point in the shape representation of the i-th individual's structure. We assume the observations X_i to be derived from the individual shape U_i. Similarly, we consider hidden random variables, i.e., V, W, etc., representing shapes for the other groups. To enable intra-group and inter-group statistical analysis, we ensure that all shape models lie in the same space by enforcing the same number of points T in all shape models.

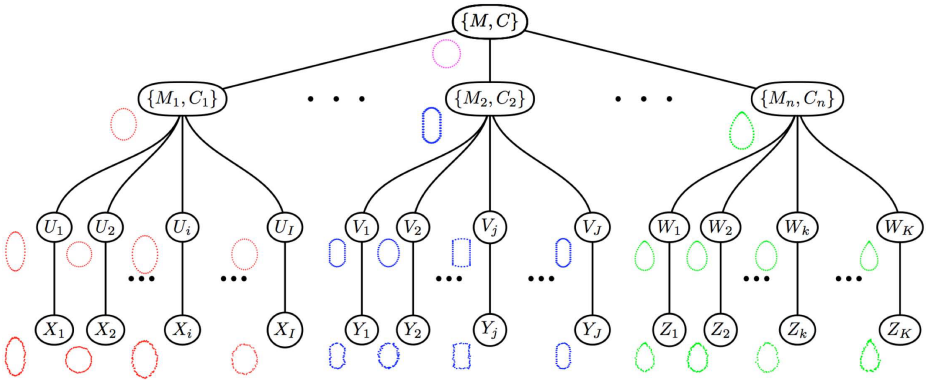


Fig. 1. Proposed Hierarchical Generative Statistical Model for Multigroup Shape Data

Group Shape Variables: Consider the first group of shapes U to be derived from a shape PDF having a mean shape M_1 and a shape covariance C_1. Consider other groups of shapes modeled analogously, i.e., V derived from a group with shape mean and covariance (M_2, C_2), W derived from a group with shape mean and covariance (M_n, C_n), etc. This paper treats the group means, i.e., M_1, M_2,$\cdots$,M_n, as hidden random variables and the group covariances, i.e., C_1, C_2,$\cdots$,C_n, as parameters. The proposed method can be generalized to treat the group covariances as random variables.

Population Shape Variables: Consider all group shape means, i.e., M_1, M_2,$\cdots$, M_n, to be derived from a single population of shapes with mean M and covariance C. In this paper, without loss of generality, we only consider two groups ($n = 2$) for simplicity.

Joint PDF: We model the joint PDF with (i) parameters M, C, C_1, C_2, (ii) group shape variables M_1, M_2, (iii) individual shape variables U, V, and (iv) data X, Y as:

$$P(M_1, M_2, U, V, X, Y|M, C, C_1, C_2) := \tag{1}$$
$$P(M_1|M, C)P(M_2|M, C)\Pi_{i=1}^{I}P(U_i|M_1, C_1)P(X_i|U_i)\Pi_{j=1}^{J}P(V_j|M_2, C_2)P(Y_j|V_j).$$

PDF of Individual Data given Individual Shape: We model $P(X_i|U_i)$, $P(Y_j|V_j)$ using current distance. Between pointsets $A := \{a_i\}_{i=1}^{I}$ and $B := \{b_j\}_{j=1}^{J}$, the squared current distance is $d_K^2(A, B) := \sum_{i=1}^{I}\sum_{i'=1}^{I} K(a_i, a_{i'}) + \sum_{j=1}^{J}\sum_{j'=1}^{J} K(b_j, b_{j'}) - 2\sum_{i=1}^{I}\sum_{j=1}^{J} K(a_i, b_j)$, where $K(\cdot, \cdot)$ is a Mercer kernel. In this paper, $K(\cdot, \cdot)$ is the Gaussian kernel with isotropic covariance $\sigma^2\mathbf{I_D}$. We use the current distance to define $P(X_i|U_i) := (1/\gamma)\exp\left(-d_K^2(X_i, U_i)\right)$, over finite support, where γ is the normalization constant. The current-distance model allows the number of points in the shape models U_i to be different from the number of boundary points in the data X_i.

Group Shape PDF: We model $P(U_i|M_1, C_1)$ as Gaussian with mean M_1 and covariance C_1 and $P(V_j|M_2, C_2)$ as Gaussian with mean M_2 and covariance C_2.

PDF of Group Variables given Population Parameters: We model $P(M_1|M, C)$ and $P(M_2|M, C)$ as Gaussian with mean M and covariance C; we choose the Gaussian

(i) to be maximally non-committal during model design and (ii) as the conjugate prior for the Gaussian means M_1, M_2. Under the Gaussian model, strange-looking shapes can be avoided by preventing over-regularization of the covariance estimate and preventing very large deviations from the mean (which are rare events under the Gaussian). More importantly, the hierarchical model alleviates this issue by producing covariance estimates that are more compact and restrict variation over fewer modes (see Figure 3).

3 Fitting the Shape Model to Data Using Monte-Carlo EM

This section presents the EM algorithm for the model-fitting optimization problem. The parameters in our model are: (i) the population mean M and covariance C and (ii) the group covariances C_1, C_2. Denoting $\theta := \{M, C, C_1, C_2\}$, the optimal model fit is:

$$\arg\max_{\theta} P(x, y|\theta) = \arg\max_{\theta} \int P(u, v, m_1, m_2, x, y|\theta) du dv dm_1 dm_2. \quad (2)$$

3.1 E Step: Sampling in Shape Space by Adapting the Hamiltonian Monte Carlo

In the i-th iteration, with parameter estimate $\widehat{\theta^i}$, the E step constructs the Q function as

$$Q(\theta|\widehat{\theta^i}) := E_{P(U,V,M_1,M_2|x,y,\widehat{\theta^i})} \log P(U, V, M_1, M_2, x, y|\theta). \quad (3)$$

Because of the analytical intractability of this expectation, we approximate $Q(\theta|\widehat{\theta^i}) \doteq \widehat{Q}(\theta|\widehat{\theta^i}) := \sum_{s=1}^{S} (1/S) \log P(u^s, v^s, m_1^s, m_2^s, x, y|\theta)$ using Monte-Carlo simulation. To sample the set of individual shapes u^s, v^s and the group-mean shapes m_1^s, m_2^s from $P(U, V, M_1, M_2|x, y, \widehat{\theta^i})$, we propose Gibbs sampling coupled with a novel adaptation of the HMC sampler [5]. Before describing the adapted HMC sampler, we outline the proposed shape-sampling algorithm for generating a sample of size S:

1. Set the sample index variable s to 0. Initialize the sampling algorithm with the sample point $s = 0$ denoted by $u^0 := \{u_i^0\}_{i=1}^I, v^0 := \{v_j^0\}_{j=1}^J, m_1^0, m_2^0$.
   Given sample point s, sample the $(s+1)$-th sample point as follows.
2. Initialized with u_i^s, $\forall i$ sample $u_i^{s+1} \sim P(U_i|v^s, m_1^s, m_2^s, x, y, \widehat{\theta^i})$.
3. Initialized with v_j^s, $\forall j$ sample $v_j^{s+1} \sim P(V_j|u^{s+1}, m_1^s, m_2^s, x, y, \widehat{\theta^i})$.
4. Initialized with m_1^s, sample $m_1^{s+1} \sim P(M_1|u^{s+1}, v^{s+1}, m_2^s, x, y, \widehat{\theta^i})$.
5. Initialized with m_2^s, sample $m_2^{s+1} \sim P(M_2|u^{s+1}, v^{s+1}, m_1^{s+1}, x, y, \widehat{\theta^i})$.
6. If $s + 1 = S$, then stop; otherwise increment s by 1 and repeat the previous 4 steps.

We ensure the independence of samples between Gibbs iteration s and the next $s + 1$, by running the HMC algorithm sufficiently long and discarding the first few samples s.

HMC is a Markov-chain Monte-Carlo sampling algorithm. HMC exploits the gradient of the log PDF for fast exploration of the space of the random variables. The HMC approach first augments the original random variables with auxiliary momentum variables, then defines a Hamiltonian function combining the original and auxiliary variables, and, subsequently, alternates between simple updates for the auxiliary variables

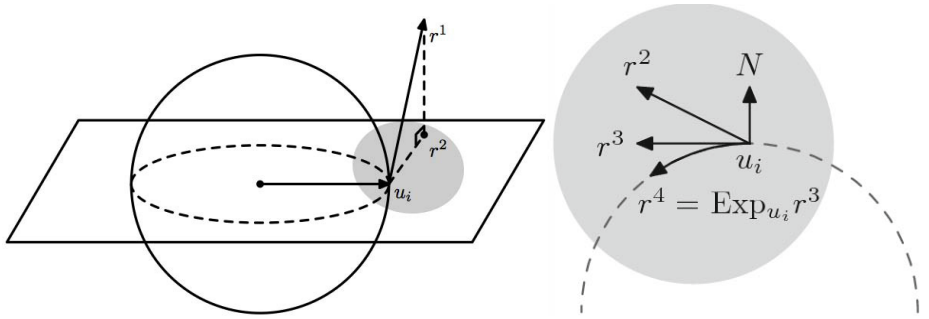


Fig. 2. Gradient Projection within HMC for Sampling in Shape Space **Left:** shows Kendall's pre-shape space [9] (dotted hypersphere) that is the intersection of the (bold) hypersphere of fixed radius ρ (i.e., $\sum_t \|u_i(t)\|^2 = \rho^2$; fixes scale) and the hyperplane through the origin (i.e., $\sum_t u_i(t) = \mathbf{0}$; fixes translation). For a pointset u_i, log-posterior gradients r^1 are projected onto the hyperplane to produce r^2 that eliminates translation. **Right:** To remove changes in scale, the resulting projection r^2 is then projected onto the tangent space at u_i, tangent to the pre-shape space, and the resulting tangent-space projection r^3 is mapped to the pre-shape space via the manifold exponential map to give r^4. The text describes the last part of the projection.

and Metropolis updates for the original variables. HMC proposes new states by computing a trajectory according to the Hamiltonian dynamics implemented with a leapfrog method and guarantees the new proposal states to be accepted with high probability. In our case, HMC requires gradients of $\log P(U, V, M_1, M_2 | x, y, \widehat{\theta^i})$ with respect to the hidden variables $\{U_i\}_{i=1}^I, \{V_j\}_{j=1}^J, M_1, M_2$.

Using HMC naively leads to pointset updates that can change the location, scale, and pose of the pointset, thereby making the sampler very inefficient. For this problem, we propose to modify HMC by replacing the gradient of the log posterior by a *projected gradient* that restricts the updated shape to Kendall shape space. As shown in Figure 2, starting with pointset u_i, the log-posterior gradient r^1 is first projected onto the preshape space to produce r^4 that has the same centroid and scale as u_i. Then, to remove rotation effects, the resulting pre-shape r^4 is rotationally aligned with the u_i, yielding r^5 (not shown in figure). These steps project the log-posterior gradient at u_i, within HMC, to generate an updated shape r^5 as part of the trajectory within HMC.

3.2 M Step: Parameter Estimation

In iteration i of the EM optimization, the M step maximizes $\widehat{Q}(\theta | \widehat{\theta^i})$ over θ and sets $\widehat{\theta^{i+1}} \leftarrow \arg\max_\theta \sum_{s=1}^S \log P(u^s, v^s, m_1^s, m_2^s, x, y | \theta)$. Equating the gradient of this objective function to zero gives closed-form optimal estimates for all parameters.

4 Group Comparison Using Permutation Testing

After the model is fit to the data, we can perform hypothesis testing to compare any pair of groups; the null hypothesis is that the two groups of data were drawn from the same

PDF. Since the shape PDF in each group is modeled using Mahalanobis distances based on means M_1, M_2 and covariances C_1, C_2, we use Hotelling's two-sample T^2 statistic to measure dissimilarity between any pair of groups. However, in 3D medical image data, the dimensionality TD can be very high compared to the number of individuals. Low sample sizes can render the F-distribution unusable. Simulating shapes with sample sizes higher than the dimensionality TD can be computationally expensive. Thus, we propose to employ distribution-free hypothesis testing, namely, permutation testing, using Hotelling's T^2 as the test statistic. Permutation testing is conservative in rejecting the null hypothesis and enhances robustness to specific modeling choices, e.g., the cardinality of the shape-model pointsets and internal model free parameters.

5 Results and Discussion

This section shows results on simulated and real data. We assume the input images, undergoing shape analysis, to be binary or soft masks, having intensities in the range $[-1, 1]$, that segment the image into the object of interest and the background. For hypothesis testing, it is less interesting to compare the performances of methods when the two groups are (i) exactly similar or (ii) differ extremely. The real challenge is in being able to reject the null hypothesis when the two groups differ in subtle ways.

For the proposed hierarchical model, we initialize the pointsets that model shape as follows. First, we solve a groupwise registration problem on the mask images, using a similarity transform, to (i) register the images, representing shape, to a common space and (ii) find an average (mask) image in that space. We assume the data to be the set of voxels on the zero crossing of the mask images warped to the common space. Then, we (i) threshold the average mask to get an object boundary, (ii) embed it as the zero level set of a signed distance-transform image, and (iii) generate a 3D triangular mesh for the zero level set using [13]. Finally, we use this mesh-vertex pointset as the *initial* value for M, m_1^0, m_2^0, $\{u_i^0\}_{i=1}^I$, and $\{v_j^0\}_{j=1}^J$. We set C, C_1, C_2 to (scaled) identity. We set the σ for the Gaussian kernel, underlying the current distance, to be the average edge length in the mesh. With this initialization, we compare the proposed method with a state-of-the-art algorithm [1], implemented in the open-source software ShapeWorks [15].

5.1 Validation on Simulated 3D Shapes

We simulate 2 groups of ellipsoidal shapes (ellipsoids in canonical form; 20 pointsets per group), where the groups are subtly different from each other. Two of the axes have length 1. The lengths of the third axis for the (i) 1st group are drawn from a Gaussian with mean 0.9 and variance 0.01 and (ii) 2nd group are drawn from a Gaussian with mean 1.1 and variance 0.01. The pointsets are then rescaled to constant norm.

The proposed method as well as ShapeWorks (i) both employ $T = 64$ points per pointset for shape modeling and (ii) both take as input equivalent information, i.e., while ShapeWorks takes as input a signed-distance-transform image (Figure 3) representing the ellipsoids implicitly, the proposed method takes as input the corresponding zero-crossing image. With $T = 64$, the average distance between a point and its nearest neighbor, in the shape pointset, is around 10 voxels. For both methods, the covariance

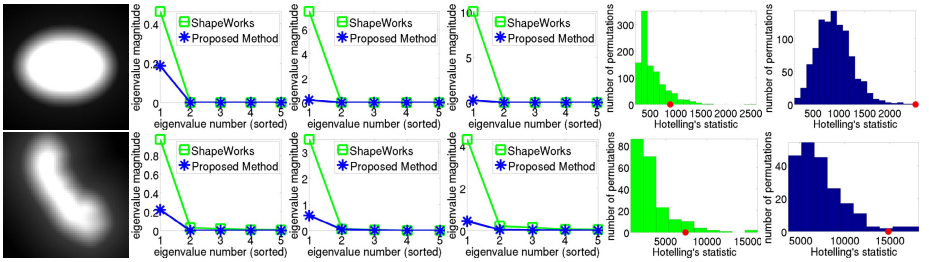


Fig. 3. Results. Top Row: *Ellipsoidal Shapes in Simulated Data.* Bottom Row: *Hippocampal Shapes in Clinical Brain MR images.* Left to Right: Distance transform image data (2D slice); Eigenspectra of the population covariance C and the group covariances C_1, C_2; Permutation distribution of Hotelling's T^2 test statistic for ShapeWorks (green) and the proposed method (blue); the red circle shows the value of the test statistic for the unpermuted group labeling.

estimates are regularized by addition of a scaled identity matrix δI, where δ is a free parameter; the experiments explore the robustness of both approaches to changes in δ.

Figure 3 (Top Row) shows the results from the proposed method compared to ShapeWorks, for the regularization parameter δ set to 10^{-4}. The proposed method leads to a fitted model that has smaller variances, at the group level as well as the population level. This indicates that the proposed method leads to a model that is more compact and fits the data better; this stems from improvements in optimal point placement and estimation of correspondences/parametrization. For the permutation distribution of the Hotelling's T^2 statistic, the p value for ShapeWorks is 0.05 and that for the proposed method is 0.001. Varying δ over $10^{-3}, 10^{-4}, \cdots, 10^{-10}$, we find that the p value for the proposed method stays at 0.001, but the p value of ShapeWorks varies and is never lower than 0.05. These results were unchanged when value of the current-distance parameter σ was multiplied by factors $\in [0.5, 2]$. This indicates that, compared to ShapeWorks, the proposed method was more robust to changes in δ and consistently produces a p value that tends to (correctly) reject the null hypothesis significantly more strongly.

5.2 Results on Clinical Brain MR Images: Hippocampal Shapes in Dementia

This section employs clinical brain magnetic resonance (MR) images from the OASIS [11] dataset. We use 10 randomly selected OASIS brains that uniformly sample the age span, including 4 cases with very-mild to mild Alzheimer's dementia and 6 controls, having hippocampus segmentations manually performed by a radiologist.

The proposed method and ShapeWorks, both, employ $T = 128$ points per pointset; the average distance between a point and its nearest neighbor is around 5 voxels. Figure 3 (Bottom Row) shows the results using $\delta = 10^{-4}$. These results were unchanged when value of the current-distance parameter σ was multiplied by factors $\in [0.5, 2]$. The proposed method leads to a fitted model that has smaller variances, indicating a compact better-fitting model. The p value for ShapeWorks is 0.07. The p value for the proposed method is 0.03 that indicates a relatively stronger rejection of the null hypothesis.

Discussion: The results show that the proposed hierarchical model and unified-optimization approach leads to compact-fitting shape models that can differentiate

subtle variations in hippocampal shapes (open-access data) better than the state of the art (open-source software). The main originality in the paper is in being able to solve the three problems of point placement, correspondence, and model-parameter estimation (given data from one or more groups) as a single optimization problem. Another key originality is in being able to sample in Kendall shape space, using a novel adaptation of HMC sampling using restricted gradients. The proposed framework can benefit from more accurate and efficient schemes for modeling and estimation.

Acknowledgments. We thank support from NIH-NCRR Center for Integrative Biomedical Computing P41-RR12553, NIH-NCBC NAMIC U54-EB005149, Royal Academy of Engineering 1314RECI076, NIH 5R01EB007688, and NSF CAREER 1054057.

References

1. Cates, J., Fletcher, P.T., Styner, M., Hazlett, H.C., Whitaker, R.: Particle-based shape analysis of multi-object complexes. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part I. LNCS, vol. 5241, pp. 477–485. Springer, Heidelberg (2008)
2. Cootes, T., Taylor, C., Cooper, D., Graham, J.: Active shape models - their training and application. Comp. Vis. Image Understanding 61(1), 38–59 (1995)
3. Coughlan, J., Ferreira, S.: Finding deformable shapes using loopy belief propagation. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part III. LNCS, vol. 2352, pp. 453–468. Springer, Heidelberg (2002)
4. Davies, R., Twining, C., Cootes, T., Taylor, C.: A minimum description length approach to statistical shape modeling. IEEE Trans. Med. Imag. 21(5), 525–537 (2002)
5. Duane, S., Kennedy, A., Pendleton, B., Roweth, D.: Hybrid Monte Carlo. Physics Letters 195, 216–222 (1987)
6. Goodall, C.: Procrustes methods in the statistical analysis of shape. J. Royal Stat. Soc. 53(2), 285–339 (1991)
7. Gu, L., Kanade, T.: A generative shape regularization model for robust face alignment. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 413–426. Springer, Heidelberg (2008)
8. Joshi, S., Kommaraji, R., Phillips, J., Venkatasubramanian, S.: Comparing distributions and shapes using the kernel distance. In: ACM Symp. Comp. Geom., pp. 47–56 (2011)
9. Kendall, D.: Shape manifolds, Procrustean metrics, and complex projective spaces. Bull. London Math. Soc. 16(2), 81–121 (1984)
10. Kotcheff, A., Taylor, C.: Automatic construction of eigen shape models by direct optimization. Medical Image Analysis 2(4), 303–314 (1998)
11. Marcus, D., Wang, T., Parker, J., Csernansky, J., Morris, J., Buckner, R.: Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neuro. 19(9), 1498–1507 (2007)
12. Neumann, A.: Graphical Gaussian shape models and their application to image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 25(3), 316–329 (2003)
13. Persson, P., Strang, G.: A simple mesh generator in Matlab. SIAM Rev. 46(2), 329–345 (2004)
14. Rangarajan, A., Coughlan, J., Yuille, A.: A Bayesian network framework for relational shape matching. In: Int. Conf. Comp. Vis., pp. 671–678 (2003)
15. SCI Institute: ShapeWorks: An open-source tool for constructing compact statistical point-based models of ensembles of similar shapes that does not rely on specific surface parameterization (2013), `http://www.sci.utah.edu/software/shapeworks.html`
16. Thodberg, H.: MDL shape and appearance models. Info. Proc. Med. Imag. 2, 251–260 (2003)
17. Vaillant, M., Glaunes, J.: Surface matching via currents. Info. Proc. Med. Imag. 19, 381–392 (2005)

Depth-Based Shape-Analysis

Yi Hong[1], Yi Gao[3], Marc Niethammer[1,2], and Sylvain Bouix[4]

[1] University of North Carolina (UNC) at Chapel Hill, NC, USA
[2] Biomedical Research Imaging Center, UNC-Chapel Hill, NC, USA
[3] University of Alabama at Birmingham, Birmingham, AL, USA
[4] Psychiatry Neuroimaging Laboratory, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

Abstract. In this paper we propose a new method for shape analysis based on the depth-ordering of shapes. We use this depth-ordering to non-parametrically define depth with respect to a normal control population. This allows us to quantify differences with respect to "normality". We combine this approach with a permutation test allowing it to test for localized shape differences. The method is evaluated on a synthetically generated striatum dataset as well as on a real caudate dataset.

1 Introduction

Population-based shape analysis is of high importance to discriminate for example normal subjects from subjects with a particular disease. Many methods for shape analysis exist. They can be subdivided into global and local analysis methods. Global analysis methods are designed to detect whether population shape differences exist [5], but cannot generally locate where these shape differences may be, which limits their ability to provide intuitive insights into the underlying biological mechanism. The main attraction of such methods is that they often avoid establishing dense correspondences between shapes through registration. In contrast, while point-to-point correspondences between shapes allow precise local shape analysis, establishing these correspondences is one of the main sources of inaccuracy as any misregistration may create artifacts with respect to the final shape analysis results. Nevertheless, a variety of methods for local shape analysis have been proposed and successfully used [8,7,1]. In this work, we explore an alternative method that allows for localized shape analysis, but only needs very limited (e.g., rigid or affine) spatial alignment of shapes. Our method uses a depth-ordering of shapes to allow to compare shape populations.

Our Main Contributions in This Paper Are:

1) We propose using depth-ordering on shapes for statistical shape analysis.
2) We develop an algorithm for the fast computation of band-depth for shapes represented through *binary* indicator functions.
3) We define statistical tests to differentiate shape populations globally and locally without an explicit computation of dense correspondences.
4) We demonstrate the method on synthetic and real datasets.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 17–24, 2014.

Sec. 2 describes how to depth-order shapes and discusses how to compute such a depth-ordering fast. Sec. 3 proposes statistical approaches using depth-ordering for shape analysis. In Sec. 4, we present experimental results for synthetic and real datasets. Sec. 5 concludes the paper with a summary.

2 Depth-Ordering of Shapes

A challenge in shape analysis is that there is no canonical ordering of shapes. Here we leverage the work on ordering of functions from the statistics literature and extend it to shapes [6]. Once defined the ordering can be used to generalize traditional order statistics, such as the median or the inter-quartile range, to shape ensembles. Recently, band-depth has been proposed as one possible way for ordering functions [9]. Intuitively, the deeper a function is buried within a dataset the more central it is. The deepest function corresponds to the within-sample median function. Band-depth has been used to define a functional boxplot [9]. It has also been extended to contour boxplots [11] defining band-depth on contours for the visualization of ensemble data. What makes band-depth attractive for shape-ordering is that shapes can be analyzed as *functions* if they are represented by indicator functions, i.e., by binary functions that are 1 inside and 0 outside of a shape [3]. Band-depth for binary shape representations relates to set unions and intersections, and it is a *natural* functional representation of shape.

Given a set of shapes as binary functions, $\{\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_n\}$, with dimension of (s_x, s_y, s_z), we vectorize them and obtain binary vectors $\mathrm{y}_i \in \{0,1\}^p$, where $p = s_x \times s_y \times s_z$. The band-depth for each shape y is defined as follows:

$$BD_n^{(j)}(\mathrm{y}) = \frac{1}{C} \sum_{1 \le i_1 < i_2 < \cdots < i_j \le n} I\{G(\mathrm{y}) \subseteq B(\mathrm{y}_{i_1}, \cdots, \mathrm{y}_{i_j})\}. \tag{1}$$

Here, $1 \le j \le J$, and J is the number of observations used for defining the band, C is a normalization constant equal to the number of admissible permutations. $G(\mathrm{y})$ is the graph of the function, $G(\mathrm{y}) = \{(\mathbf{x}, \mathrm{y}(\mathbf{x})) : \mathbf{x} \in \mathbb{I}\}$. B is the band delimited by the observations given as its arguments. That is, $B(\mathrm{y}_{i_1}, \cdots, \mathrm{y}_{i_j}) = \{(\mathbf{x}, \mathrm{y}(\mathbf{x})) : \mathbf{x} \in \mathbb{I}, min_{r=i_1,\cdots,i_j}\, \mathrm{y}_r(\mathbf{x}) \le \mathrm{y}(\mathbf{x}) \le max_{r=i_1,\cdots,i_j}\, \mathrm{y}_r(\mathbf{x})\}$. $I\{.\}$ denotes the indicator function, which evaluates to 1 if the graph of the function is within the band, or to 0, otherwise. Since the band depth on binary functions may result in many ties for the resulting depth, it can be modified [9] to

$$MBD_n^{(j)}(\mathrm{y}) = \frac{1}{C} \sum_{1 \le i_1 < i_2 < \ldots < i_j \le n} \lambda_m\{A(\mathrm{y}; \mathrm{y}_{i_1}, ..., \mathrm{y}_{i_j})\} \tag{2}$$

where $A_j(\mathrm{y}) \equiv A(\mathrm{y}; \mathrm{y}_{i_1}, ..., \mathrm{y}_{i_j})$ and $A_j(y) \equiv \{\mathbf{x} \in \mathbb{I} : min_{r=i_1,...,i_j} \mathrm{y}_r(\mathbf{x}) \le \mathrm{y}(\mathbf{x}) \le max_{r=i_1,...,i_j} \mathrm{y}_r(\mathbf{x})\}$, m is the observation's dimension, $\lambda_m(\mathrm{y}) = \lambda(A_j(\mathrm{y}))/\lambda(\mathbb{I})$ and λ is the Lebesgue measure on $\mathbb{R}^m$.

However, albeit its conceptual simplicity, one of the main limitations of the band-depth computation is its computational complexity. Therefore, recently a

fast method to compute band-depth has been proposed [10] which is based on computing local curve ranks. However, the proposed algorithm is ill-suited for binary shape representations as it does not consider ranking ties for the modified band-depth (MBD) nor special cases where curves can change ordering without affecting the rank of a specific curve. For binary representations ranking can be avoided as at any point only two values are possible. The computation of MBD can then be accomplished efficiently for $J = 2$. Our algorithm is as follows:

Step 0) Given n binary volumes, $\{\mathrm{y}_i\}_{i=1}^n$, vectorize them: $\mathrm{y}_i \in \{0,1\}^p$.
Step 1) At each location, k, for a given value $v(k) \in \{0,1\}$, we count the number of functions that have a value larger (n_a), smaller (n_b) or equal (n_t) to v:
- if $v(k) = 0$, then $n_a = \sum_{i=1}^n \mathrm{y}_i(k)$, $n_b = 0$, and $n_t = n - n_a - 1$
- if $v(k) = 1$, then $n_a = 0$, $n_b = \sum_{i=1}^n (1 - \mathrm{y}_i(k))$, and $n_t = n - n_b - 1$

Step 2) We then calculate the number of pairwise combinations containing $v(k)$:

$$C_k(v(k)) = n_a n_b + (n_a + n_b)n_t + n_t(n_t - 1)/2 + (n_a + n_b + n_t).$$

For binary functions n_a and n_b cannot simultaneously be different from zero. Furthermore, $n_a + n_b + n_t = n - 1$, which simplifies the expression to

$$C_k(v(k)) = (n_a + n_b)n_t + n_t(n_t - 1)/2 + n - 1.$$

Step 3) The modified band-depth for a curve y_i is then

$$MBD(\mathrm{y}_i) = \frac{1}{p}\binom{n}{2}^{-1} \sum_{k=1}^{p} C_k(\mathrm{y}_i(k)),$$

where the notation $C_k(\mathrm{y}_i(k))$ denotes computing C_k based on the coefficients n_a, n_b, n_t given by the value of y_i at location k.

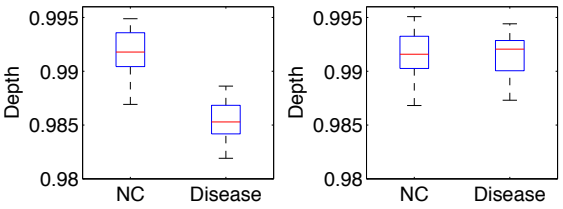
In comparison to the original band-depth algorithm our computing complexity is reduced from $O(pn^3)$ to $O(pn)$. This makes the computation for large populations and large multi-dimensional shapes (we will focus on 2D surfaces in 3D here) possible. Furthermore, it enables us to perform permutation tests based on band-depth computations as discussed in Sec. 3.

3 Statistics Using Depth-Ordering

Band-depth measures the relationship between a shape and a reference population. A higher value indicates the shape is closer to the median, and a lower one indicates the shape is a potential outlier with respect to the reference population. Based on this property of band-depth we can perform global shape analysis as described in Sec. 3.1 as well as local shape analysis as described in Sec. 3.2. For all these analyses we assume that shapes have been pre-aligned as appropriate. Typically this will either involve a rigid, similarity or affine alignment of shapes to a template or some form of unbiased atlas-building method. The choice of transform will depend on the objective of a given study. E.g., if size differences should be included rigid alignment would be appropriate. The key ingredient to performing statistics using depth-ordering is to compute depth-ordering with respect to a reference population of shapes that are used as a non-parametric model of shapes, with respect to which depth is measured.

3.1 Global Shape Analysis

Given a dataset $\{R_i\}$ containing a reference population of shapes, we compute the band-depth for a given datum D, from a set of input test shapes $\{D_j\}$, by computing the band-depth for all the data in $\{R_i\} \bigcup D$ and assign the resulting band-depth for D to D, denoted as $BD(D; \{R_i\})$. This is substantially different from directly computing the band-depths for the dataset $\{D_j\}$. In the proposed method the reference population forms a "yard-stick" by which to judge data-depth. In the latter case data-depth is defined with respect to the dataset itself which is problematic as band-depth does not have a sense of directionality, but only a sense of how close a data-element is to the deepest data-element. For different populations which should be discriminated this consequentially leads to a data-mingling which no longer allows for a discrimination of the populations. To illustrate this effect Fig.1 shows results for the two approaches for the synthetic striatum data described in detail in Sec. 4.1. The proposed approach can clearly differentiate the populations whereas a joint computation of the band-depth is not successful.

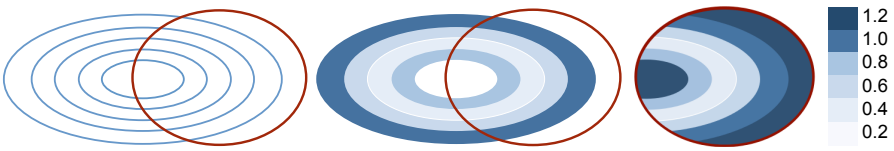


(a) w.r.t. ref. (b) without ref.

Fig. 1. Global shape analysis using band-depth with (a) and without (b) a reference population. A reference population allows detection of shape differences.

3.2 Local Shape Analysis

The local analysis is based on the the central regions of the reference population. By gradually adding the deepest shapes according to their band-depth, one can assign α values, the proportion of the added reference shapes, describing the "centrality" of a shape population at each point in the domain. A test shape can be overlaid on this centrality map and the corresponding α values recorded on its surface, thus providing a local measure of shape abnormality. Fig. 2 illustrates this concept for a population of two-dimensional shapes. Given the reference shapes shown in Fig. 2(a), we compute their α-central level sets based on the band-depth. As shown in Fig. 2(b), the deepest shape has the lowest α value (light blue) and the most outlying shape has the highest α value (dark blue). A local measure of "belonging" to the population can then be computed for a test shape by tracing the α-central region it traverses as shown in Fig. 2(c). Note that some regions of the shape may not be covered by the reference shape population, so we use a dilation procedure starting from the boundary of $\alpha = 1$ central region, evolving at a constant speed until all voxels of the volume are covered, e.g., the regions colored with the darkest blue in Fig. 2(c).



(a) Reference and test shapes (b) α-central region levelsets (c) Local measure of shape

Fig. 2. (a) Reference shape population (blue contours) defines (b) α-central level sets (light to dark blue corresponds to most to least central) that provides a local measure (c) of how deeply a shape (red) is buried with respect to the reference population. The dilated region is colored with the darkest blue and has a value greater than 1.

4 Experimental Results

4.1 Synthetic Data Experiments

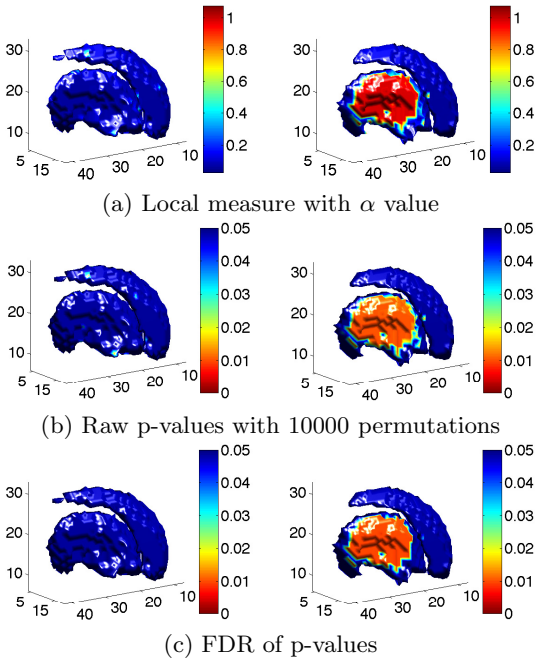
Using synthetic data allows us to introduce a predefined shape change which we wish to recover using our proposed approach. We used the technique described in [2] to generate large data sets of realistic shapes with known deformations. In short, a manifold learning technique is used to generate arbitrarily many shapes from a small training sample. A joint clustering algorithm is then applied to parcellate each shape's surface into small regions which are consistently located across all shapes. Finally, a Log-Euclidean framework is used to introduce smooth, invertible and anatomically realistic deformations to one or multiple regions as defined by the clustering. For this application, we generated 160 shapes based on 27 manually traced striatums. We then modified 80 of them by thickening the putamen. We evenly divide 80 normal controls into two groups. One is used for the reference group (NC-Train), and the other is for testing (NC-Test). In the 80 abnormal subjects, we randomly pick 40 of them for testing.

Global Analysis. To test for group separability, we performed a permutation test (10000 permutations) on the mean depth of the NC-Test versus the abnormal group. When using the NC-Train to compute band depth, the resulting p-value is 0, indicating the normal controls and the disease subjects are significantly different. On the other hand, as shown in Fig. 1, when pooling all shapes together to compute their band depths, no significant difference is detected.

Local Analysis. We used the NC-Train group to estimate the set of α central regions of a "normal" population and tested the median shapes of the NC-Test and abnormal population against it as shown in Fig. 3(a). In addition, to displaying α values, we performed a non parametric statistical analysis, based on a permutation test procedure. A template median shape is first computed, then α regions are computed for each group and differences in α values are recorded at each point of the template. 10000 permutations are performed and we count the number of α values that are larger than the one computed with no permutation. The p-values are shown in Fig. 3(b). Fig. 3(c) reveals the false discovery rate (FDR) of p-values.

4.2 Real Data Experiments

Magnetic Resonance Images (MRI) of the brains of 28 neuroleptic-naïve female subjects diagnosed with Schizotypal Personality Disorder (SPD) and of 25 female normal control subjects were acquired on a 1.5-T General Electric MR scanner. Spoiled-gradient recalled acquisition (SPGR) images (voxel dimensions 0.9375 0.9375 1.5 mm) were obtained coronally. The caudate nucleus was delineated manually by an expert. This data set was used in previous volumetric and shape analysis studies [4]. All the caudate shapes are pre-aligned using rigid transformation.



(a) Local measure with α value

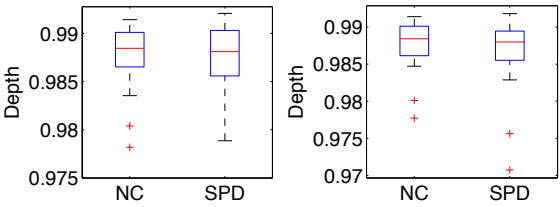
(b) Raw p-values with 10000 permutations

(c) FDR of p-values

Fig. 3. Local analysis on synthetic striatum with α values (a) on normal (left) and abnormal (right) median shapes, and corresponding p-values (b), as well as the FDR of p-values (c)

Global Analysis. Unlike our synthetic data experiment, we do not have enough controls to have non overlapping training and testing data sets. We thus use a leave-one-out method to compute the depth for normal controls, and use the whole control group as the reference to compute the depth for the SPD group. Fig. 4 demonstrates the global differences between normal controls and SPDs, for both left and right caudate. We also use 10000 permutation tests to measure the significant difference of depth for both normal controls and SPD group, resulting in p-values, 0.48 for the left caudate and 0.21 for the right caudate. This indicates based on the global depth-based analysis, both left and right caudates are not significantly different in the SPD and the NC populations.

Local Analysis. For the local shape analysis, we compute the local α values for the median shape of the SPD group, by using the normal control group as the reference shape population. Similar to the permutation test for the synthetic data, the



(a) Left caudate (b) Right caudate

Fig. 4. Band-depth for left and right caudate

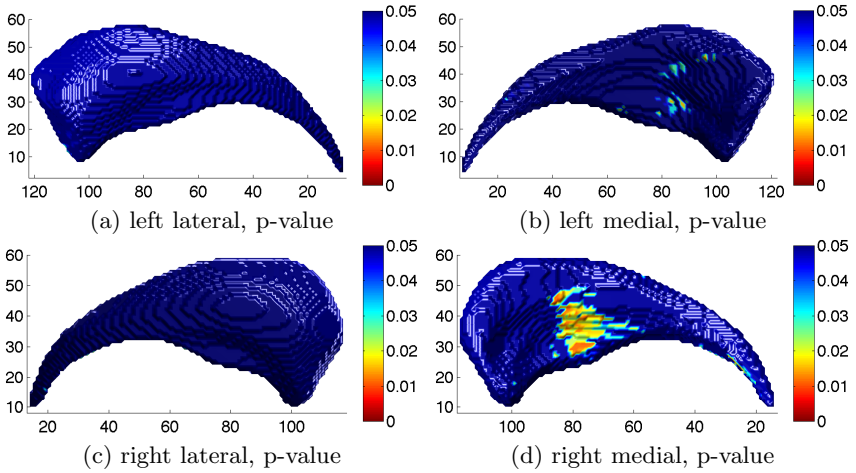


Fig. 5. Local p-values. Left caudate SPD median with respect to NC (top), right caudate SPD median with respect to NC (bottom).

p-values with 10000 permutations are shown in Fig. 5. Our method can capture the abnormal region, but based on the local p-values only relatively small regions of the SPD group seem to be significantly different from normal controls.

5 Discussion and Conclusion

In this paper we presented a shape analysis framework that can provide both global and local information, yet does not require complex processes to establish point-to-point correspondences. Instead we use the notion of band-depth of functions to order shapes according to how well they "fit in" a shape ensemble. This method allows for the definition of a median and α-central regions of a population, which can then be used to compare different population of shapes.

Different from [3], which focuses on augmenting a population atlas with statistical information using weighted band depth, we proposed a fast algorithm to compute the band-depth of shapes represented by binary maps, and most importantly showed how band depth can be used to provide both global and local statistical tests to differentiate between populations. In contrast to other deformation based tools for shape analysis, our approach is non-parametric and naturally captures the probability of a shape belonging to a population. Although it does not provide physical measurements of displacement, these can be computed by deformation or a distance transform to the population median.

Our method was successfully tested with synthetic data, where we were able to clearly separate groups and localize an artificially induced shape change. In addition, our real data experiment supports previous results on shape differences in the caudate of females with SPD, namely, a right sided shape difference in the body of the caudate.

Acknowledgements This work was supported by NIH grants R01-MH082918, P41EB002025, and R01-HL105241.

References

1. Davies, R., Twining, C., Taylor, C.: Statistical Models of Shape: Optimisation and Evaluation. Springer, London (2008) 17
2. Gao, Y., Bouix, S.: Synthesis of realistic subcortical anatomy with known surface deformations. In: Levine, J.A., Paulsen, R.R., Zhang, Y. (eds.) MeshMed 2012. LNCS, vol. 7599, pp. 80–88. Springer, Heidelberg (2012) 21
3. Hong, Y., Davis, B., Marron, J.S., Kwitt, R., Niethammer, M.: Weighted functional boxplot with application to statistical atlas construction. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part III. LNCS, vol. 8151, pp. 584–591. Springer, Heidelberg (2013) 18, 23
4. Levitt, J.J., Styner, M., Niethammer, M., Bouix, S., Koo, M.S., Voglmaier, M.M., Dickey, C.C., Niznikiewicz, M.A., Kikinis, R., McCarley, R.W., Shenton, M.E.: Shape abnormalities of caudate nucleus in schizotypal personality disorder. Schizophrenia Research 110(1-3), 127–139 (2009) 22
5. Loncaric, S.: A survey of shape analysis techniques. Pattern Recognition 31(8), 983–1001 (1998) 17
6. López-Pintado, S., Romo, J.: On the concept of depth for functional data. Journal of the American Statistical Association 104, 718–734 (2009) 18
7. Miller, M.I.: Computational anatomy: Shape, growth, and atrophy comparison via diffeomorphisms. NeuroImage 23, S19–S33 (2004) 17
8. Styner, M., Lieberman, J.A., Pantazis, D., Gerig, G.: Boundary and medial shape analysis of the hippocampus in schizophrenia. Medical Image Analysis 8(3), 197–203 (2004) 17
9. Sun, Y., Genton, M.: Functional boxplots. Journal of Computational and Graphical Statistics 20, 316–334 (2011) 18
10. Sun, Y., Genton, M.G., Nychka, D.W.: Exact fast computation of band depth for large functional datasets: How quickly can one million curves be ranked? Stat. 1, 68–74 (2012) 19
11. Whitaker, R.T., Mirzargar, M., Kirby, R.M.: Contour boxplots: A method for characterizing uncertainty in feature sets from simulation ensembles. IEEE Transactions on Visualization and Computer Graphics 19(12), 2713–2722 (2013) 18

Genus-One Surface Registration via Teichmüller Extremal Mapping

Ka Chun Lam[1], Xianfeng Gu[2], Lok Ming Lui[1]

[1] Department of Mathematics, The Chinese University of Hong Kong
[2] Department of Computer Science, State University of New York at Stony Brook
{kclam,lmlui}@math.cuhk.edu.hk, gu@cs.sunysb.edu

Abstract. This paper presents a novel algorithm to obtain landmark-based genus-1 surface registration via a special class of quasi-conformal maps called the Teichmüller maps. Registering shapes with important features is an important process in medical imaging. However, it is challenging to obtain a unique and bijective genus-1 surface matching that satisfies the prescribed landmark constraints. In addition, as suggested by [11], conformal transformation provides the most natural way to describe the deformation or growth of anatomical structures. This motivates us to look for the unique mapping between genus-1 surfaces that matches the features while minimizing the maximal conformality distortion. Existence and uniqueness of such optimal diffeomorphism is theoretically guaranteed and is called the Teichmüller extremal mapping. In this work, we propose an iterative algorithm, called the Quasi-conformal (QC) iteration, to find the Teichmüller extremal mapping between the covering spaces of genus-1 surfaces. By representing the set of diffeomorphisms using Beltrami coefficients (BCs), we look for an optimal BC which corresponds to our desired diffeomorphism that matches prescribed features and satisfies the periodic boundary condition on the covering space. Numerical experiments show that our proposed algorithm is efficient and stable for registering genus-1 surfaces even with large amount of landmarks. We have also applied the algorithm on registering vertebral bones with prescribed feature curves, which demonstrates the usefulness of the proposed algorithm.

1 Introduction

Surface registration is increasingly used in morphometric analysis. By finding a meaningful one-to-one correspondence between anatomical surfaces, statistical shape analysis, processing of signals on anatomical surfaces (e.g., denoising or filtering) and age-related comparison can be achieved. In landmark-based registrations, landmarks are extracted to guide the registration process to obtain a meaningful transformation. Through labeling landmarks, medical experts and doctors can get involved in the process to assure good correspondences between the surfaces. However, obtaining a unique and bijective registration that matches features consistently is generally challenging, especially when a large number of landmark constraints are enforced. Developing an effective algorithm for registration is therefore very important.

Surface registration between simple surfaces, such as simply-connected open surfaces or genus-0 closed surfaces, has been extensively studied. However, as far as we

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 25–32, 2014.

know, very few literatures have been reported on the registration between genus-one surfaces. The high-genus topology of the surfaces poses a great challenge to register the surfaces. For example, the vertebral shape is commonly analyzed through simple geometric measurements of dimensions, which only describe limited features of the complex vertebral shape. In order to provide a more comprehensive description, a more sophisticated landmark-based surface registration is essential for analyzing both local and global geometric information of a vertebral shape.

Motivated by this, we are interested in searching for the unique and bijective landmark-matching diffeomorphism which minimizes the maximal conformality distortion. The conformality distortion measures how far the mapping is deviated from a conformal mapping, and hence it measures the local geometric distortion.. The existence and uniqueness of such a mapping is theoretically guaranteed by the Quasi-conformal Teichmüller theory [2], and is named the Teichmüller extremal map. In this paper, we propose a novel algorithm to compute the Teichmüller extremal map between genus-1 surfaces. Experiments on vertebral bones are also reported to show the accuracy and effectiveness of the proposed algorithm.

2 Previous Work

Landmark-based registration has been widely studied in medical imaging, computer graphics and computer visions. Various algorithms have been proposed to match feature landmarks consistently. For example, Bookstein et al.[1] proposed to obtain a registration that matches landmarks as much as possible using a thin-plate spline regularization (or biharmonic regularization). Gu et al. [4,5] proposed to compute the conformal parameterizations of human brain surfaces for registration using harmonic energy minimization and holomorphic 1-forms. Conformal registration is advantageous for the preservation of the local geometry. However, it cannot align landmark features, such as sulci landmarks on brain surfaces, consistently. Sometimes, deformation between objects might not be conformal. Instead, certain amount of angular distortion could be introduced. To tackle with this problem, quasi-conformal mappings have been applied to obtain surface registration with bounded conformality distortion [8,9]. Introduction of time-dependent vector fields for registration is also proposed [7,3]. For example, Glaunés et al. in [3] presented to generate large deformation diffeomorphisms of a sphere, with given displacements of a finite set of template landmarks. The time dependent vector fields facilitate the optimization procedure, but the computational cost of the algorithm is comparatively more expensive.

3 Mathematical Background

3.1 Quasi-Conformal Map

Quasi-conformal maps are orientation preserving homeomorphisms between Riemann surfaces with bounded conformality distortion. Intuitively, they take infinitesimal

circles to infinitesimal ellipses of bounded eccentricity. Mathematically, $f : \mathbb{C} \to \mathbb{C}$ is quasi-conformal provided that it satisfies the Beltrami equation:

$$\frac{\partial f}{\partial \bar{z}} = \mu(z)\frac{\partial f}{\partial z} \tag{1}$$

for some complex-valued function μ satisfying $\|\mu\|_\infty < 1$. The function μ is a measure of non-conformality and is named the *Beltrami coefficient*. In particular, a map f is conformal at p if $\mu(p) = 0$. Denote $i = \sqrt{-1}$ and $f = u + iv$. From the Beltrami equation (1),

$$\mu(f) = \frac{(u_x - v_y) + i(v_x + u_y)}{(u_x + v_y) + i(v_x - u_y)} \tag{2}$$

Let $\mu(f) = \rho + i\tau$. We have the following linear combinations between u_x, u_y, v_x and v_y:

$$\begin{cases} -v_y = \alpha_1 u_x + \alpha_2 u_y \\ v_x = \alpha_2 u_x + \alpha_3 u_y \end{cases}; \quad \begin{cases} -u_y = \alpha_1 v_x + \alpha_2 v_y \\ u_x = \alpha_2 v_x + \alpha_3 v_y \end{cases} \tag{3}$$

where $\alpha_1 = \frac{(\rho-1)^2+r^2}{1-\rho^2-r^2}$; $\alpha_2 = \frac{2r}{1-\rho^2-r^2}$; $\alpha_3 = \frac{1+2\rho+\rho^2+r^2}{1-\rho^2-r^2}$. By taking divergence on both sides of equations (3), we obtain

$$\nabla \cdot \left(A \begin{pmatrix} u_x \\ u_y \end{pmatrix} \right) = 0; \quad \nabla \cdot \left(A \begin{pmatrix} v_x \\ v_y \end{pmatrix} \right) = 0, \quad \text{where } A = \begin{pmatrix} \alpha_1 & \alpha_2 \\ \alpha_2 & \alpha_3 \end{pmatrix} \tag{4}$$

According to the Quasi-conformal Teichmüller theory, a quasi-conformal map can be uniquely determined up to Mobiüs transformations. Ambiguity of the Mobius transformation can be eliminated by providing three points correspondence, in which a unique solution can be obtain from equation (4).

3.2 Teichmüller Extremal Map

Let $\mu(f)$ be the Beltrami coefficient of f. Define the maximal dilation of f to be:

$$K(f) = \frac{1 + \|\mu(f)\|_\infty}{1 - \|\mu(f)\|_\infty}. \tag{5}$$

Using maximal dilation, we can define extremal map as:

Definition 1 *Let $f : S_1 \to S_2$ be a quasi-conformal mapping between S_1 and S_2. f is said to be an extremal mapping if for any quasi-conformal mapping $h : S_1 \to S_2$ isotopic to f relative to the boundary,*

$$K(f) \leq K(h) \tag{6}$$

It is uniquely extremal if the inequality (6) is strict when $h \neq f$.

Another kind of mapping, called the *Teichmüller mapping*, is closely related to the extremal mapping. Teichmüller mapping is defined as follows:

Definition 2 *Let $f : S_1 \to S_2$ be a quasi-conformal mapping. f is said to be a Teichmüller mapping associated to the quadratic differential $q = \varphi dz^2$ where $\varphi : S_1 \to \mathbb{C}$ is a holomorphic function if its associated Beltrami differential is of the form:*

$$\mu(f) = k\frac{\overline{\varphi}}{|\varphi|} \tag{7}$$

for some constant $k < 1$ and quadratic differential $q \neq 0$ with $||q||_1 = \int_{S_1} |\varphi| < \infty$.

Let S_1 and S_2 be Riemann surfaces with the same topology. Let $\{p_i\}_{i=1}^n \in S_1$ and $\{q_i\}_{i=1}^n \in S_2$ be the corresponding interior landmark constraints. A Teichmüller mapping f between S_1 and S_2, which satisfies the landmark constraints, is actually the unique extremal map. With this, both uniqueness and existence of landmark matching Teichmüller extremal map can be guaranteed. We can therefore obtain a unique landmark matching registration by searching for an optimal Beltrami coefficient whose maximal dilatation is the minimum. For details, please refer to [2].

4 Proposed Algorithms

In this section, we explain our algorithm for obtaining a feature aligned Teichmüller extremal mapping between genus-1 surfaces. The basic idea is to first embed the surfaces into their universal covering spaces and find the Teichmüller extremal mapping between their conformal parameterizations.

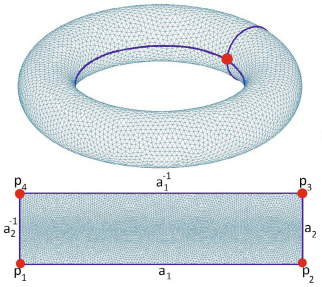


Fig. 1. Torus & Ω

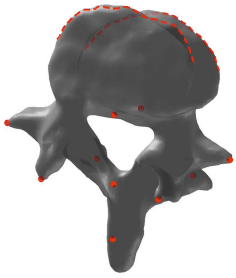


Fig. 2. Vertebral bone 1

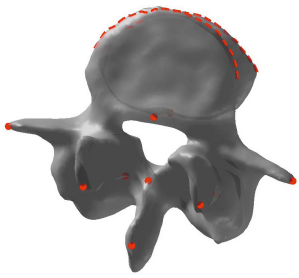


Fig. 3. Vertebral bone 2

4.1 Embed Genus-One Surface into the Euclidean Plane

The embedding of the genus-1 surface is computed using the Ricci flow method introduced by Gu et al. [6]. The basic idea of Ricci flow is to conformally deform the metric $g = (g_{ij}(t))$ according to its induced Gaussian curvature $K(t)$. Mathematically, we have

$$\frac{dg_{ij}(t)}{dt} = -2(K(t) - \bar{K})g_{ij}(t) \tag{8}$$

where we set $\bar{K} = 0$ for genus-one to be the target curvature. Convergence of this process is guaranteed by Hamilton's theorem. $g(\infty)$ is the desired uniformization metric.

Let S be a genus-1 surface and p be a base point for S. Two closed loops based at p are introduced to slice the genus-1 surface into the fundamental domain. With the uniformization metric, the fundamental domain can be conformally embedded onto a 2D domain $\Omega \in \mathbb{R}^2$, called the fundamental polygon (See Figure 1). Denote the boundaries and vertices of the polygon as $\{a_1, a_2, a_1^{-1}, a_2^{-1}\}$ and $\{p_i\}$ respectively. The boundary pairs $\{a_1, a_1^{-1}\}$,$\{a_2, a_2^{-1}\}$ and vertices $\{p_i\}$ correspond to the closed loop and the single based point introduced. Note that a_i and a_i^{-1}, $i = 1, 2$ are related by $\varphi_i(a_i) = a_i^{-1}$, where φ_i are translations in $\mathbb{R}^2$. Therefore, periodic constraints are enforced in the boundaries of the fundamental polygon. With this conformal parameterization, registration can be done on the fundamental domains instead of the complex genus-1 surfaces. For details, please refer to [6,10].

4.2 Computing the Teichmüller Extremal Mapping between Parameter Domains

Let Ω_1 and Ω_2 be the fundamental polygons of two genus-1 surfaces S_1 and S_2 respectively. Denote the boundaries and vertices of Ω_1 and Ω_2 by $\{a_1, a_2, a_1^{-1}, a_2^{-1}\}$, $\{p_i^{S_1}\}$ and $\{b_1, b_2, b_1^{-1}, b_2^{-1}\}$, $\{p_i^{S_2}\}$ respectively. As the boundary cuts of S_1 and S_2 may not be consistent, only periodic constraints are considered during the registration. Let $\{r_k\}_{k=1}^n$ and $\{q_k\}_{k=1}^n$ be the landmark correspondences on Ω_1 and Ω_2 respectively. Mathematically, the problem of finding Teichmüller extremal mapping between the fundamental domains can be formulated as follows:

$$f = \mathbf{argmin}_{f:\Omega_1 \to \Omega_2} \|\mu(f)\|_\infty \tag{9}$$

subject to:

- $\mu(f) = k\frac{\bar{\varphi}}{|\varphi|}$ where $0 \leq k < 1$ and $\varphi : \Omega_1 \to \mathbb{C}$ is integrable holomorphic;
- $\varphi_i(f(a_i)) = f(a_i^{-1})$ for $i = 1, 2$; (Periodic constraints) (10)
- $f(p_i^{S_1}) = p_i^{S_2}$ for $i = 1, ..., 4$; (Base points consistency) (11)
- $f(r_k) = q_k$ for $k = 1, 2, ..., n$. (Landmark constraints) (12)

To solve the above minimization problem, we propose an iterative scheme called the Quasi-conformal (QC) iteration. The basic idea is to find a path in the space of all Beltrami coefficients, which approaches from $\mu = 0$ to the unique admissible Beltrami coefficient ν^* of Teichmüller type. The process is summarized in Algorithm 1. For the convergence of Algorithm 1, please refer to [12].

5 Experimental Results

To evaluate the proposed algorithm, we apply it on the vertebral bones to compute the Teichmüller extremal map between 5 pairs of vertebral bones with prescribed feature points and landmark curves as landmarks (See Figure 2 & 3). There are two landmark curves labeled on the top and bottom side of the cortical rim and ten features marked on other parts of each vertebral bone. To register between a pair of vertebral bones, we first parameterize them into the fundamental domains Ω_1 and Ω_2 by the Ricci flow

Algorithm 1: QC-iteration

Input: Ω_1 and Ω_2; landmark constraints $\{r_k\}$ and $\{q_k\}$.
Output: Optimal Beltrami coefficient ν and the Teichmüller extremal map f

Initial: $\nu_0 = 0$;
repeat
  Update $f = (u, v)$ by solving (4) with ν_{n+1} and constraints (10),(11) and (12);
  Compute $\mu_{n+1} := \mathcal{A}(\mathcal{L}(\nu_n))$, where $\mathcal{L}$ is the laplace smoothing operator and

$$\mathcal{A}(\mu) = \int_{\Omega_1} |\mu| d\Omega_1 \Big/ \int_{\Omega_1} d\Omega_1;$$

  Update $f = (u, v)$ by solving (4) with μ_{n+1} and constraints (10),(11) and (12);
  Set $\nu_{n+1} := \mu(f)$, where $\mu(f)$ is the Beltrami coefficients of f
until $\|\nu_{n+1} - \nu_n\|_\infty \leq \epsilon$;

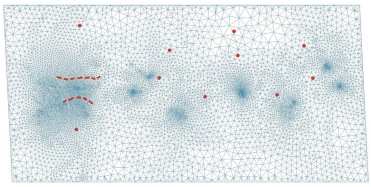


Fig. 4. Fundamental polygon Ω_1

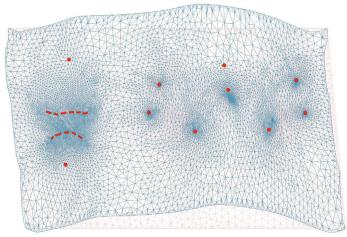


Fig. 5. Registered polygon and Ω_2

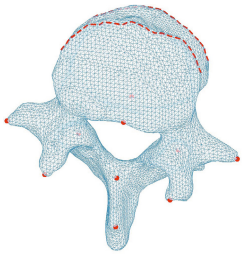


Fig. 6. Vertebral bone S_1

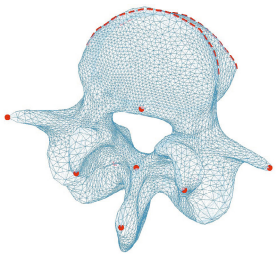


Fig. 7. Resultant registration

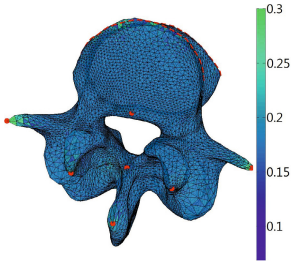


Fig. 8. $|\mu|$ on surface

method. Using the QC iteration, the Teichmüller extremal mapping $f : \Omega_1 \to \Omega_2$ which satisfies the landmark constraints is obtained. Since no hard constraints is enforced on the cutting boundaries in the algorithm, the cutting boundaries of Ω_1 can move freely on the universal covering space, which satisfy the periodic conditions. Figure 5 shows the obtained Teichmüller extremal map between the covering spaces. Once the Teichmüller extremal map is computed, we can obtain the registration between the vertebral bones S_1 and S_2 by a composition of functions $\phi_2^{-1} \circ f \circ \phi_1 = T : S_1 \to S_2$. The resultant registration is shown in Figure 7. The mesh is obtained by deforming the source vertebral bone (Figure 6) to the target surface (Figure 3). Landmark curves and

Table 1. Summary of the comparison experiment

Method	$e_{\max}$	e_{mean}	$\|\mu\|_\infty$	$SD(\|\mu\|)$	d_H	Time (s)
Proposed	0	0	0.4193	0.0147	0.82	10.91s
rigid ICP	0.1467	0.0389	$1.08e^{-13}$	$5.33e^{-15}$	12.44	4.46s
non-rigid ICP	0.0798	0.0402	0.9841	0.1710	4.55	223.07s

feature points are exactly matched after the registration process. Figure 8 also shows $|\mu(f)|$ of the Teichmüller Extremal mapping, which is represented by the color on the vertebral bone surface. An even color distribution on the surface and a small standard deviation of the BC norm of 0.001823 indicate that the resultant mapping is actually of Teichmüller type. By the properties of Teichmüller map, the registration obtained is guaranteed to be bijective. This demonstrates that our proposed algorithm can effectively provide the unique registration result which minimizes the maximal conformality distortion. We have also computed the Teichmüller extremal mappings between a set of vertebral bones to construct the mean surface (Figure 9). Both feature points and the landmark curves are well-preserved, illustrating that the landmarks are consistently matched under the proposed registration algorithm.

To validate the invariance of the choice of cutting boundaries during the embedding process, we manually labeled two arbitrary simple closed loops (blue-red loops in Figure 10) with the same base point and run the proposed algorithm. Figure 11 shows the histogram of the optimal Teichmüller type BCs $|\mu_g|$ and $|\mu_{br}|$ from the cases of green loops and blue-red loops respectively. Experiment shows that both registration results are coincident, with $\||\mu_{\text{g}}| - |\mu_{\text{br}}|\|_\infty = 0.0021$, indicating that our proposed algorithm is invariant to the initial choice of the cutting boundaries. We have also compared our implementation with rigid ICP and non-rigid ICP. The result is summarized in Table 1. For ease of comparison, we first normalize every vertebral bone to fit into a unit cube. In terms of the mean and maximum landmark matching errors (e_{mean}, $e_{\max}$), our proposed method outperforms the two point-based registration methods. The Hausdorff distance d_H between the registration result and the target also shows that our proposed method has a better overlay percentage to the target object. With the sacrifice of the registration accuracy, almost no conformality distortion is introduced by the rigid ICP, while the non-rigid ICP produces a large distortion of 0.9841. Our proposed algorithm thus provides a balance between the computation requirement and the registration accuracy.

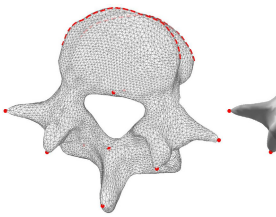


Fig. 9. Vertebral bone mean surface

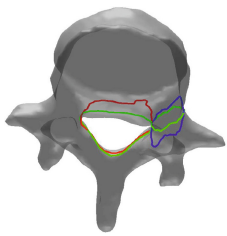


Fig. 10. Different loops

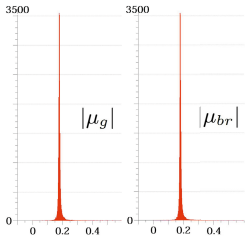


Fig. 11. $|\mu_g|$ & $|\mu_{br}|$

6 Conclusion and Future Works

This paper presents a novel method to compute the Teichmüller extremal mapping with prescribed landmark correspondences between genus-1 surfaces, which minimizes the maximal conformality distortion. By the Teichmüller theory, existence and uniqueness of such mapping is guaranteed. We applied the proposed algorithm for the vertebral bone registration and the construction of mean surface of vertebral bones. Experimental results show that our method is effective in computing bijective feature aligned registration with smallest maximal conformality distortion. In the future, we plan to extend the proposed method to higher-genus surfaces and apply the method to more real applications in medical imaging for disease analysis.

Acknowledgements. This research is supported by the HKRGC GRF (CUHK Project ID: 404612).

References

1. Bookstein, F.L.: Principal Warps: Thin-Plate splines and the decomposition of deformations. IEEE Trans. Pattern Anal. Machine Intell. 11(6), 567–585 (1989)
2. Gardiner, F., Lakic, N.: Quasiconformal Teichmuller Theory. American Mathematics Society (2000)
3. Glaunès, J., Vaillant, M., Miller, M.I.: Landmark Matching via Large Deformation Diffeomorphisms on the Sphere. JMIV 20(1-2), 179–200 (2004)
4. Gu, X.F., Wang, Y., Chan, T.F., Thompson, P.M., Yau, S.T.: Genus zero surface conformal mapping and its application to brain surface mapping. IEEE Trans. Med. Imag. 23(8), 949–958 (2004)
5. Hurdal, M.K., Stephenson, K.: Discrete conformal methods for cortical brain flattening. Neuroimage 45(1), 86–98 (2009)
6. Jin, M., Kim, J., Luo, F., Gu, X.F.: Discrete surface Ricci flow. IEEE Trans. Visual. Comput. Graphics 14(5), 1030–1043 (2008)
7. Joshi, S., Miller, M.I.: Landmark matching via large deformation diffeomorphisms. IEEE Trans. Image Processing 9(8), 1357–1370 (2000)
8. Lui, L.M., Thiruvenkadam, S., Wang, Y., Chan, T.F., Thompson, P.M.: Optimized conformal parameterization of cortical surfaces using shape based matching of landmark curves. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part I. LNCS, vol. 5241, pp. 494–501. Springer, Heidelberg (2008)
9. Lui, L.M., Wang, Y., Chan, T.F., Thompson, P.M.: Landmark constrained genus zero surface conformal mapping and its application to brain mapping research. Appl. Numer. Math. 57(5), 847–858 (2007)
10. Lui, L.M., Wen, C.F.: Geometric Registration of High-Genus Surfaces. SIIMS 7(1), 337–365 (2014)
11. Thompson, D.W.: On growth and form (1942)
12. Lui, L.M., Gu, X.F., Yau, S.T.: Convergence analysis of an iterative algorithm for Teichmüller maps via harmonic energy optimization. Math. Comp. (2014)

Subject-Specific Prediction Using Nonlinear Population Modeling: Application to Early Brain Maturation from DTI

Neda Sadeghi[1], P. Thomas Fletcher[1], Marcel Prastawa[2], John H. Gilmore[3], and Guido Gerig[1]

[1] Scientific Computing and Imaging Institute, University of Utah
[2] GE Global Research
[3] Department of Psychiatry, University of North Carolina

Abstract. The term *prediction* implies expected outcome in the future, often based on a model and statistical inference. Longitudinal imaging studies offer the possibility to model temporal change trajectories of anatomy across populations of subjects. In the spirit of subject-specific analysis, such normative models can then be used to compare data from new subjects to the norm and to study progression of disease or to predict outcome. This paper follows a statistical inference approach and presents a framework for prediction of future observations based on past measurements and population statistics. We describe prediction in the context of nonlinear mixed effects modeling (NLME) where the full reference population's statistics (estimated fixed effects, variance-covariance of random effects, variance of noise) is used along with the individual's available observations to predict its trajectory. The proposed methodology is generic in regard to application domains. Here, we demonstrate analysis of early infant brain maturation from longitudinal DTI with up to three time points. Growth as observed in DTI-derived scalar invariants is modeled with a parametric function, its parameters being input to NLME population modeling. Trajectories of new subject's data are estimated when using no observation, only the first or the first two time points. Leave-one-out experiments result in statistics on differences between actual and predicted observations. We also simulate a clinical scenario of prediction on multiple categories, where trajectories predicted from multiple models are classified based on maximum likelihood criteria.

1 Introduction

Longitudinal data analysis can provide further insight into growth, degeneration or disease progression by analyzing change trajectories rather than snapshots in time. In this setting, individual subjects' trajectories can be compared to the normative models computed via population modeling. One can then identify the timing of deviation from typical trajectories, interventions can be targeted toward a specific developmental period, or predicted trajectories can be used for assessment of disease risk during prodromal stage or for measuring efficacy of disease-modifying therapies, for example.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 33–40, 2014.

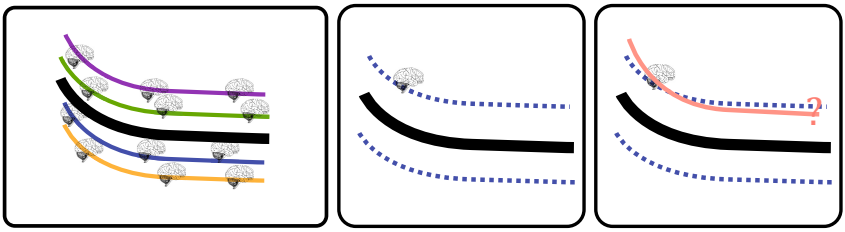


Fig. 1. Prediction based on reference population and new individual's scan(s). Left: Population trajectory constructed based on the reference population. Middle: Population trajectory along with predicted interval. Right: Can we predict the new individual's trajectory based on new scan(s) and the reference population?

The term *prediction* can be used in very different contexts and with different goals. E.g., genetics may predict risk for disease, a patient score or disease status may be predicted from imaging biomarkers, or physiological age is predicted from sets of measurements. Here, we focus on the notion of statistical inference to predict a future observation of a new subject given a model for the temporal trajectory, comprehensive statistics from training on population data, and a set of past observations from this subject (Fig. 1). The predicted observation with confidence bounds, or more general the prediction of the whole trajectory with variability, can then be used to estimate deviation from the norm. In addition, given normative models for multiple groups, for example for different patient categories and/or controls, one can derive prediction trajectories for each group and classify based on the most likely category.

Prediction of an individual trajectory is possible even if not all the observations for all time points would be available for that subject by pooling the data from other subjects in the study along with the available observations for the individual. Analysis of longitudinal data needs to take into account the correlation due to repeated measures, variability between subjects, often unbalanced spacing due to acquisitions at different time points and missing data. All these favor the use of mixed effects models, which represent a class of statistical methods that model the correlation of measurements of an individual along with modeling the mean response of a population over time.

The proposed methodology is generic with respect to any type of data. Here, we demonstrate proof of concept with a clinical infant neuroimaging study. Longitudinal brain imaging is increasingly used in clinical studies as it provides a superior characterization of developmental trajectories compared to cross-sectional studies [2,3]. Such studies have mostly focused on population analysis. However, individuals would likely benefit from subject-specific assessments, comparing an individual's image-derived data at each given age to the norm, and predictions of subject-specific growth trajectories and intervals based on measurements of only one or two time points, predictions which may improve early detection and therapeutic intervention.

Key aspects discussed in this paper are the selection of optimal nonlinear models to characterize temporal trajectories, building of normative models for populations, and the development of a statistical inference framework to predict future observations based on new scans' data and statistics from a reference population (Sec. 2). Predicted trajectories from multiple groups can then be used for classification and the results are presented in (Sec. 3).

2 Method

Reference Population. To prepare discussion of the prediction scheme, we briefly describe the nonlinear mixed effects model (NLME) [4]. In the mixed effects model, the observed data are assumed to be a combination of both fixed effects, β, parameters associated with the entire population (or at least within a subpopulation), and random effects, b, that are specific to an individual drawn at random from the population. Random effects account for the heterogeneity that is present in the population as these effects vary among subjects. NLME is a generalization of linear mixed effects and nonlinear regression, some or all of the fixed or random effects enter the model nonlinearly. Each individual's response is modeled as:

$$y_i = f(\phi_i, t_i) + e_i, \tag{1}$$

where $\phi = A_i\beta + B_i b_i$, and b_i are random effects with distribution $\mathcal{N} \sim (0, \Psi)$. A_i and B_i are design matrices that indicate whether a specific fixed or random effect should be included in the model. The function f can be any nonlinear function, and e_i is the measurement error and is assumed normally distributed $\mathcal{N} \sim (0, \sigma^2)$. Random effects and measurement errors are assumed to be independent.

The likelihood function for the mixed effects model is written as:

$$L(\beta, \Psi, \sigma^2|y) = \prod_{i=1}^{M} p(y_i|\beta, \Psi, \sigma^2). \tag{2}$$

Since non-observable random effects are part of the model, we must integrate out random effects; thus, the marginal density of y_i becomes:

$$p(y_i|\beta, \Psi, \sigma^2) = \int p(y_i|\beta, b_i, \sigma^2) p(b_i|\Psi, \sigma^2) db_i. \tag{3}$$

The population growth parameters β and variance components Ψ and σ^2 are estimated by maximizing the likelihood equation of (3). In general there is no closed form solution to equation 3. We approximate the integral in (3) using Taylor expansion of the model function f around conditional modes of random effects b and the current estimate of β [4].

The distribution of the maximum likelihood estimator β of the fixed effects based on the linear mixed effects approximation [4] is written as:

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \left[\sum_{i=1}^{M} \hat{X}_i^T \hat{V}_i^{-1} \hat{X}_i\right]^{-1}\right), \tag{4}$$

where $\hat{V} = \hat{Z}_i \Psi \hat{Z}_i^T + \sigma^2 I_{n_i}$, $\hat{X}_i = \frac{\partial f_i}{\partial \beta^T}|_{\hat{\beta}, \hat{b}_i}$, $\hat{Z}_i = \frac{\partial f_i}{\partial b_i^T}|_{\hat{\beta}, \hat{b}_i}$.

Choice of Nonlinear Function. The choice of the nonlinear function,f, is to be seen as study and data specific. Nonlinear models provide a more parsimonious model compared to their linear counterparts (i.e. polynomials). More important, as extrapolation of data beyond the observed range plays an important role in the prediction of future values, model parameters of nonlinear schemes tend to have natural interpretations [5]. Data analyzed here are radial diffusivity (RD) measurements of diffusion tensor imaging (DTI) of early brain development and age is the covariate. White matter is known to mature more rapidly in the first year of life than second, with continued maturation but at a much slower rate into adulthood [1,6]. This favors functions that have asymptotic behavior such as exponential, logistic, or Gompertz. We used Akaike Information Criterion (AIC) [7] for model selection, where $AIC = -2log(L_i) + 2n_{par}$, L_i is the likelihood of model i and n_{par} is the number of model parameters. Among the tested models, the Gompertz function provided the lowest AIC measure, so that this function was chosen for infant DTI modeling. Using the Gompertz function, response y is modeled as $y = \alpha e^{-d\, e^{-rt}}$, where α is the final asymptotic value, parameter r specifies the decay in the growth rate, and parameter d controls the difference between the final and initial values of y. All the parameters of the Gompertz function are used as fixed effects. Parameters α and d were chosen as random effects as they provided the best model fit as measured by lower AIC.

Prediction of New Individual Trajectory. Upon availability of new data for an individual, we can use the reference population parameters along with the individual's available data to predict a personalized growth trajectory. We substitute $\hat{\beta}$,$\hat{\psi}$ and $\hat{\sigma}$ of the reference population for the unknown parameters to predict an approximate empirical Bayes' estimate of b_i. Once the subject's random effects are estimated, the individual's growth trajectory and future values can be predicted.

The prediction of b_i can be calculated from the posterior distribution of $p(b_i|y_i)$ using Bayes' rule:

$$p(b_i|y_i, \beta, \Psi, \sigma^2) = \frac{p(y_i|\beta, b_i, \Psi, \sigma^2)p(b_i|\Psi)}{p(y_i|\beta, \Psi, \sigma^2)}. \tag{5}$$

By maximizing the log of the posterior density of b_i, we obtain the following objective function:

$$l(b_i) = -\frac{1}{\sigma^2}(y_i - f(\beta, b_i))^T(y_i - f(\beta, b_i)) - b_i^T\Psi^{-1}b_i. \tag{6}$$

Once $\hat{b}_i$ is estimated, $E[\hat{b}_i] \simeq \hat{\Psi}\hat{Z}_i^T\hat{V}_i^{-1}(y_i - f(A_i\hat{\beta} + B_i\hat{b}_i, t_i) + \hat{Z}_i\hat{b}_i)$, we can construct continuous growth trajectories of the ith subject. The ith subject prediction for the corresponding responses y_i is: $E[\hat{y}_i|b_i] = f(x_i^T\hat{\beta} + z_i^T\hat{b}_i, t)$, where x_i represents a vector of fixed effects covariates and z_i represents a vector of covariates corresponding to random effects.

Individual's Prediction Interval. By knowing the sampling distribution $b_i \sim \mathcal{N}(\hat{b}_i, \hat{W})$ and $\hat{W} = \hat{\Psi} - \hat{\Psi}\hat{Z}_i^T V_i^{-1}\hat{Z}_i\hat{\Psi}$, we can employ a Monte Carlo simulation

to approximate the subject-specific prediction interval. One thousand samples of $\beta \sim \mathcal{N}\left(\hat{\beta}, \left[\sum_{i=1}^{M} \hat{X}_i^T \hat{V}_i^{-1} \hat{X}_i\right]^{-1}\right)$, $b \sim \mathcal{N}(\hat{b}_i, \hat{W})$, and $e \sim \mathcal{N}(0, \hat{\sigma})$ were generated from their respective distributions. Subsequently, 1000 trajectories were constructed by the NLME model. The prediction interval for the "new" subject can be calculated by constructing the $1 - \alpha$ range of values for a given time point, t_{ij}. At each t_{ij}, $\alpha/2$ and $1 - \alpha/2$ percentiles were calculated as the lower and upper limits of the subject-specific interval.

Classification of a New Individual. An individual's predicted trajectory is a combination of the estimated population parameters (fixed effects), and subject specific random effects. If multiple reference subpopulations are available, denoted as $c \in C$, mixed effects modeling as described earlier can be used to estimate fixed effects $\hat{\beta}_c$, variance-covariance of random effects $\hat{\Psi}_c$, and variance of noise $\hat{\sigma}_c$ for each subpopulation c. Upon availability of new scans for an individual, random effects and individual's subject's trajectory can be predicted as shown. Once random effects (trajectories) are predicted for the new individual, we classify the subject to belong to the subpopulation that has the highest likelihood given the individual's predicted random effects $\hat{b}_{ic}$ and reference subpopulation parameters $\hat{\Psi}_c$. We assign a subject to a subpopulation c where $p(\hat{b}_{ic}|\hat{\Psi}_c)$ has the highest value among C. This method takes into account not only the subpopulation trajectory $\hat{\beta}_c$ to predict $\hat{b}_{ic}$, but also the heterogeneity of subjects $\hat{\Psi}_c$ presented in the subpopulation for classification.

3 Validation and Results

The following discussion is based on clinical data from an ongoing infant DTI study but focuses on validation by mimicking its potential clinical use via leave-one-out experiments. We verify two aspects of the proposed methodology based on two scenarios: 1) predicting a future observation for a new individual, does it fall into the range of the norm and what is the difference between prediction and observed value, and 2) having models for two reference populations and thus two prediction trajectories for new individuals, what is the classification based on the more likely population (Fig. 4).

We have access to DTI data of 26 subjects with a total of 59 DTI scans (neonate: 23, year 1: 22, year 2: 14) from a normative study, with preprocessing and unbiased atlas mapping following [6]. The reference population trajectory is estimated using all subjects excluding the one used for testing. Fig. 2 shows the estimated subject growth trajectory along with the subject-specific prediction interval for RD values of a test subject, overlaid on the population model (gray). The trajectory of the individual is predicted as discussed in section 2.

Figure 2 left shows the predicted trajectory based on only the first time point (solid blue curve and light blue region) with the two left out measurements (red dots). Upon availability of more time points, future observations are predicted with increased precision (Fig. 2 right).

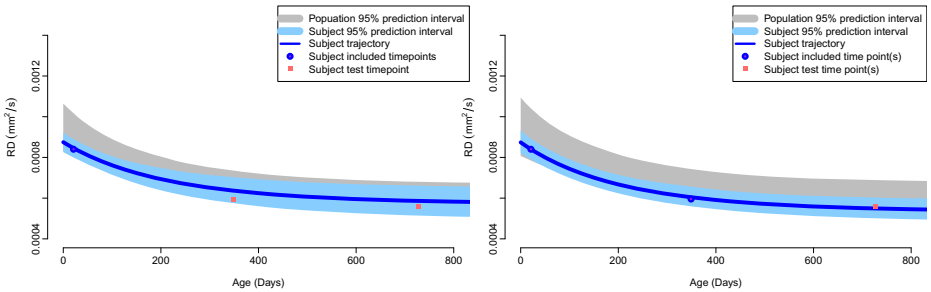


Fig. 2. Subject prediction interval compared to the overall prediction for RD of PTR. Left: subject-specific interval calculated based on only one time point (neonate). Right: subject-specific interval calculated based on scans at neonate and 1 year. Blue dots represent observations used for prediction, and red dots actual observed measurements.

Figure 3 shows the difference between observed and predicted values for RD of posterior thalamic radiation (PTR). The population trajectory was constructed using scans of 25 subjects. We then test our methodology for predicting RD at years 1 and 2, i.e. closeness of the predicted trajectory (solid blue line) to the observed measures (red dots) shown in Fig. 2. Predicting 1 year values was based on 19 subjects with at least neonate and 1 year time points, whereas prediction at 2 years was obtained from 9 subjects that have three time points available. Without any observation available for a new individual, the predicted RD value at years 1 and 2 are the population averages. However, as one observation becomes available the RD values at years 1 and 2 can be predicted with more accuracy (Fig. 3 middle). With two observations, variability of differences between predicted and observed RD values is further reduced (Fig. 3 right).

The predicted subject trajectories can also be used for classification. To illustrate a clinical scenario with two populations (e.g. controls vs. disease), we construct an example using two different regions representing two categories. Figure 4 illustrates the concept of classification into the categories PTR or splenium. We construct the population trajectories for RD of both PTR and splenium of 25 subjects, and then predict left out subject's RD trajectory for a test region as if the region could be either splenium or PTR. Looking at the test region's first time point (Fig. 4), it seems that it would be splenium, but the second time point is more similar to PTR; illustrating that analysis at single time points can easily lead to contradictory results. We predict the RD trajectories of the test region based on the PTR population (purple dashed line in Fig. 4) and splenium (yellow dashed line). With the two predictions for the test region, we then use the classification method of section 2 to assign the most likely category. The experiment is repeated for all the subjects with available scans at neonate and year 1. Overall, only one subject's splenium was misclassified as PTR and vice-versa, compared to the overlapping distributions and also conflicting classifications at single time points.

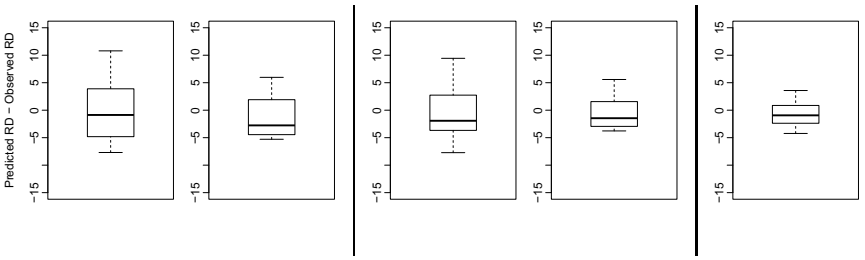


Fig. 3. Distribution of differences between observed and predicted values for radial diffusivity (RD) of PTR. Left two figures: using population mean as predicted value for prediction at year 1 (left) and year 2 (right). Middle two figures: using RD at neonate and reference population to predict RD at year 1 (left) and year 2 (right). Right figure: using population reference and both neonate and year 1 RD to predict year 2 RD. Note: the reference population was estimated based on 26 subjects with 59 scans (minus the test subject), but only 9 subjects had data for all three time points. Plots illustrate increased prediction performance from using no observation over one and then two observations (left, middle, right). RD differences displayed here have been scaled up by 10^5.

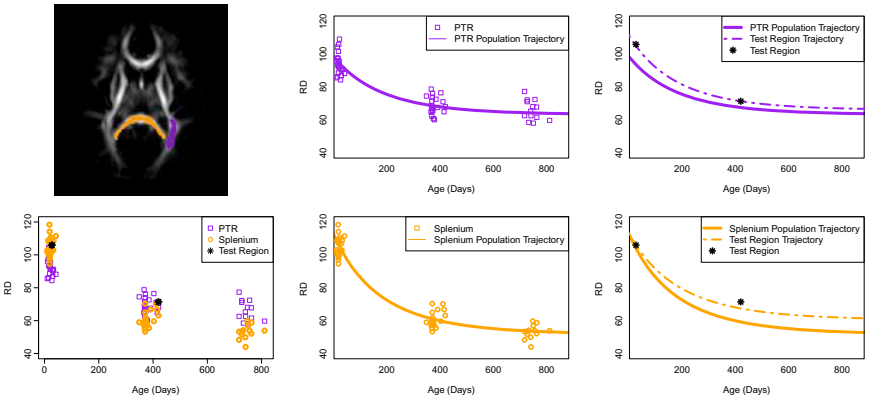


Fig. 4. Test if a new region's RD measures are representing PTR or splenium. Left: scatterplot of longitudinal measurements of radial diffusivity (RD) for the two categories. Middle: population trajectories for PTR and splenium. Right: trajectories of the test region predicted for each category (dashed lines). The test regions' trajectory with highest likelihood determines the classification, here classified into the PTR category. RD values displayed here have been scaled up by 10^5.

4 Discussion and Conclusion

This paper describes a framework for prediction based on a statistical inference approach in the mixed-effects-modeling setting, with emphasis on its application to *discrete-time data*, the use of *nonlinear parametric* temporal functions, and

employing nonlinear mixed-effects-modeling (NLME). Given a new, unseen individual, estimated subject-specific trajectories and prediction intervals not only take into account parameter estimates of the normative population but also consider existing observations from the new individual. Experimental tests with a leave-one-out scheme clearly demonstrate that the resulting subject-specific prediction interval, representing uncertainty, is steadily narrowing from using no additional observation over the use of one and then two individual measurements. Here, we are making the assumption that an individual will have a temporal trajectory similar to that of the reference population. This allows transfer of information from the estimated population model to the prediction of a new individual trajectory, thus specifically tailored to the new subject.

Although current work focuses on application to clinical longitudinal neuroimaging studies, we here decided to demonstrate the potential use of prediction-based classification in a simulation procedure. Whereas comparing trajectories of two different anatomical regions may be seen somewhat artificial in view of standard clinical studies with control and patient populations, it nevertheless highlights the interesting properties and new potential of the proposed scheme. Results clearly demonstrate that classification based on estimated trajectories with use of reference population statistics and few observations is more powerful than cross-sectional classification based on data at single time-points.

Acknowledgments. Supported by NIH grants: Conte Center MH064065, NAMIC EB005149, ACE HD055741, and CAMID NIDA DA022446-01.

References

1. Dubois, J., Dehaene-Lambertz, G., Perrin, M., Mangin, J.F., Cointepas, Y., Duchesnay, E., Le Bihan, D., Hertz-Pannier, L.: Asynchrony of the early maturation of white matter bundles in healthy infants: quantitative landmarks revealed noninvasively by diffusion tensor imaging. Hum Brain Mapp 29, 14–27 (Jan 2008)
2. Giedd, J.N., Snell, J.W., Lange, N., Rajapakse, J.C., Casey, B.J., Kozuch, P.L., Vaituzis, A.C., Vauss, Y.C., Hamburger, S.D., Kaysen, D., Rapoport, J.L.: Quantitative magnetic resonance imaging of human brain development: ages 4-18. Cereb. Cortex 6(4), 551–560 (1996)
3. Kraemer, H.C., Yesavage, J.A., Taylor, J.L., Kupfer, D.: How can we learn about developmental processes from cross-sectional studies, or can we? Am J Psychiatry 157(2), 163–171 (Feb 2000)
4. Lindstrom, M.L., Bates, D.M.: Nonlinear mixed effects models for repeated measures data. Biometrics 46, 673–687 (Sep 1990)
5. Pinheiro, J.C., Bates, D.M.: Mixed-Effects Models in S and S-Plus. Springer (2000)
6. Sadeghi, N., Prastawa, M., Fletcher, P.T., Wolff, J., Gilmore, J.H., Gerig, G.: Regional characterization of longitudinal DT-MRI to study white matter maturation of the early developing brain. Neuroimage 68, 236–247 (Mar 2013)
7. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6), 716–723 (1974)

BrainPrint: Identifying Subjects by Their Brain

Christian Wachinger[1,2], Polina Golland[1], and Martin Reuter[1,2]

[1] Computer Science and Artificial Intelligence Lab, MIT, Cambridge, US
[2] Massachusetts General Hospital, Harvard Medical School, Boston US

Abstract. Introducing *BrainPrint*, a compact and discriminative representation of anatomical structures in the brain. *BrainPrint* captures shape information of an ensemble of cortical and subcortical structures by solving the 2D and 3D Laplace-Beltrami operator on triangular (boundary) and tetrahedral (volumetric) meshes. We derive a robust classifier for this representation that identifies the subject in a new scan, based on a database of brain scans. In an example dataset containing over 3000 MRI scans, we show that *BrainPrint* captures unique information about the subject's anatomy and permits to correctly classify a scan with an accuracy of over 99.8%. All processing steps for obtaining the compact representation are fully automated making this processing framework particularly attractive for handling large datasets.

1 Introduction

Is it possible to identify an individual based on their brain? Are cortical folding patterns unique to a person, similar to a fingerprint? While the unique complexity of the brain may indicate that an unambiguous identification should be possible, there is currently little empirical research that can speak to these questions. One difficulty for identifying the subject of a given brain is that longitudinal changes caused by aging or disease may significantly alter the brain morphometry. Additionally, scanning artifacts, inhomogeneities, and different imaging protocols can cause changes in intensity values in magnetic resonance scans, further complicating the identification. Therefore, a subject-specific brain signature must be both stable across time and insensitive to imaging artifacts. Moreover, it needs to provide a holistic representation of the brain to ensure subject identification even if certain parts change. Finally, small changes in the brain should map to small changes in the representation to permit a robust identification.

Here, we introduce *BrainPrint*, a holistic representation of the brain anatomy, containing the shape information of an ensemble of cortical and subcortical structures. The inclusion of only shape information has the advantage to remain independent from the local intensity values in the scan. Moreover, the variety of the different structures included in the *BrainPrint* yields an extensive characterization of the brain anatomy. We quantify the shape information by calculating the spectrum of the Laplace-Beltrami operator (LBO) on both triangular meshes that represent boundary surfaces, e.g., the white matter surface, and tetrahedral

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 41–48, 2014.

meshes for volumetric representations of individual structures. We then derive a classifier that identifies a subject from an MRI scan based on its *BrainPrint*. We achieve robustness in the identification by letting each brain structure vote independently for the subject's identity. Not only does our classifier identify previously encountered subjects with high accuracy, but it also determines whether a query brain belongs to an unknown subject, not yet represented in the existing database.

An alternative approach to calculate the similarity between scans could be based on image registration [3,5]. However, real applications of such identification methods require large datasets and the cost for aligning a new scan to all scans in the database becomes prohibitive for a large number of scans. *BrainPrint* introduces a new framework that is especially beneficial when working with large datasets widely available today. The first step extracts information from the image, based on the segmentation of anatomical structures. The second step transfers this information into a compact and discriminative representation, the *BrainPrint*. Any further processing is conducted on this representation, which takes less memory and permits easier calculations and comparisons than the original scan.

1.1 Related Work

A 3D object can be represented by the space that it occupies (3D volume representation, e.g., voxels, tetrahedra meshes) or by representing its boundary (2D surface representation, e.g., triangle meshes). Reuter et al. [10] introduced the "shapeDNA" and demonstrated that the spectra of 3D solid objects and their 2D boundary surfaces contain complementary information: the spectra of the 2D boundary surface was capable of distinguishing two isospectral 3D solids (GWW-prisms). Therefore, we propose to combine the information from both the 3D solid and 2D boundary shape representations.

While there has been previous work analyzing the shapeDNA for single brain structures [1,9,11], to the best of our knowledge this is the first study that evaluates its application to cortical structures and a wide range of subcortical structures. Importantly, we investigate the joint modeling of the ensemble. Additionally, most prior work computes the shapeDNA for triangular surface meshes [1,8], while we also work with tetrahedral volume tessellations. Given that the Laplace spectra are isometry invariant, the 2D boundary representation alone may yield a weaker descriptor, due to the large set of potential (near-) isometric deformations. For example, a closed 2D surface with a protrusion pointing inwards yields the same descriptor as one with the protrusion pointing outwards, while the spectra of the enclosed 3D solids differ.

2 Shape Descriptor

We segment anatomical structures from brain scans with FreeSurfer [2]. Next, we compute a compact shape representation that captures important shape information and facilitates the further processing. Since image intensity varies across

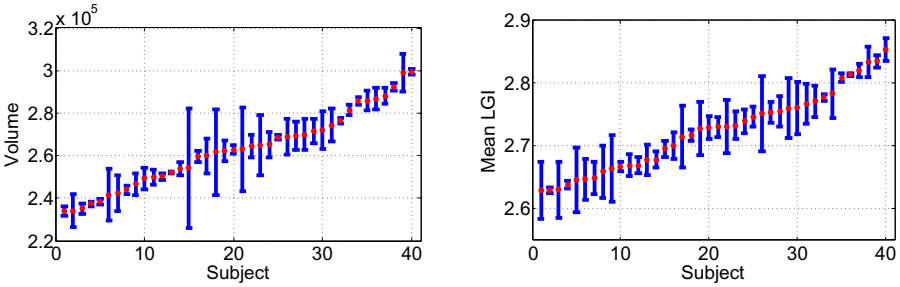


Fig. 1. Mean and standard deviation of the volume (left) and of the mean local gyrification index (right) of the cortex for 40 subjects. Statistics are calculated over several longitudinal scans per subject.

scans, we focus on geometrical properties. Example representations are the volume and the local gyrification index (LGI) of a structure. While volume will be affected by brain atrophy, quantifying the gyrification may be more robust to longitudinal changes, assuming that the folding patterns of the brain remain stable. The LGI was used previously to identify gyral abnormalities [12]. We transform this local measure into a global shape descriptor by computing the mean LGI over the surface. Fig. 1 shows the mean and standard deviation of these measures calculated from several longitudinal scans per subject. The large variance and overlap across subjects indicates that such representations are not well suited for identifying subjects.

In this work we use the shapeDNA [10] as a shape descriptor, which performed among the best in a recent comparison of methods for non-rigid 3D shape retrieval [6]. The ShapeDNA is computed from the intrinsic geometry of an object by calculating the Laplace-Beltrami spectrum. Considering the Laplace-Beltrami operator Δ, we obtain the spectrum by solving the Laplacian eigenvalue problem (Helmholtz equation) $\Delta f = -\lambda f$ using the finite element method. The solution consists of eigenvalue $\lambda_i \in \mathbb{R}$ and eigenfunction f_i pairs (sorted by eigenvalues, $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots$). To be independent of the objects' scale, we normalize the eigenvalues $\lambda' = \mathrm{vol}^{\frac{2}{D}}\lambda$, where vol is the Riemannian volume of the D-dimensional manifold (i.e., the area for 2D surfaces) [10]. The first l non-zero eigenvalues form the shapeDNA: $\boldsymbol{\lambda} = (\lambda'_1, \ldots, \lambda'_l)$.

The eigenvalues are isometry invariant with respect to the Riemannian manifold, meaning that length-preserving deformations will not change the spectrum. This important property permits the comparison of subjects by directly comparing the shapeDNA, without the need for alignment. While isometric non-congruent surfaces exist (e.g., bending a sheet of paper), two solid bodies embedded in $\mathbb{R}^3$ are isometric if and only if they are congruent (translated, rotated and mirrored). A second property is that the spectrum continuously changes with topology-preserving deformations of the object boundary. Fig. 2 illustrates the eigenfunctions of the cerebral cortex boundary. The eigenfunctions show natural vibrations of the shape when oscillating at a frequency specified by the square root of the eigenvalue.

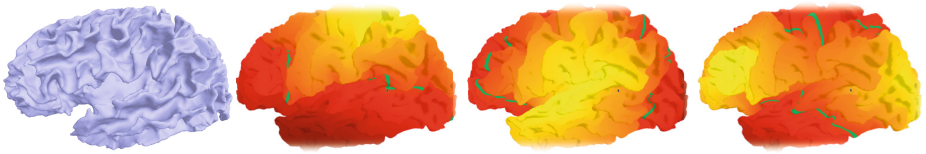


Fig. 2. Left cerebral cortex and first eigenfunctions of the LBO calculated on the surface (yellow – positive, red – negative, and green – zero)

We compute the spectra for all cortical and subcortical structures on the 2D boundary surfaces (triangle meshes) and additionally on the full 3D solid (tetrahedra meshes) for the cortical structures (white and pial surfaces in both hemispheres), forming the *BrainPrint* $\Lambda = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_\eta)$. Triangle meshes of the cortical surfaces are obtained automatically for each hemisphere using FreeSurfer. Surface meshes of subcortical structures are constructed via marching cubes from the FreeSurfer subcortical segmentation. To construct tetrahedral meshes, we remove handles from the surface meshes, uniformly resample the output to 60K vertices, and create the volumetric mesh with the gmsh package [4]. We use the linear finite element method [10] with Neumann boundary condition (zero normal derivative) to compute the spectra of the tetrahedral meshes.

3 Classifier

We derive a classifier to assign a new scan to one of the subjects in the database. Since the segmentation or tessellation of specific structures may fail in certain cases, we propose a robust classifier that handles missing information. We build a classifier by combining the results from weak classifiers operating on specific brain structures.

Assuming n subjects $\mathcal{C}_1, \ldots, \mathcal{C}_n$ and N scans in a database ($N \geq n$, for repeated scans of subjects). Each scan has its associated *BrainPrint* $\Lambda_1, \ldots, \Lambda_N$. Let $S_k \subset \{1, \ldots, N\}$ denote scans for subject $\mathcal{C}_k$. The probability that a the new scan with *BrainPrint* Λ shows subject $\mathcal{C}_k$ is

$$p(\mathcal{C}_k|\Lambda) = \frac{p(\Lambda|\mathcal{C}_k) \cdot p(\mathcal{C}_k)}{\sum_\nu p(\Lambda|\mathcal{C}_\nu) \cdot p(\mathcal{C}_\nu)} \propto \prod_{s=1,\ldots,\eta} p(\boldsymbol{\lambda}_s|\mathcal{C}_k), \tag{1}$$

where we assume a uniform class probability $p(\mathcal{C}_k) \propto 1$ and the conditional independence of structures given the subject. The likelihood is multivariate normal distributed $p(\boldsymbol{\lambda}_s|\mathcal{C}_k) \sim \mathcal{N}(\boldsymbol{\lambda}_s; \boldsymbol{\mu}_s^k, \Sigma_s)$ with the subject mean $\boldsymbol{\mu}_s^k = \frac{1}{|S_k|} \sum_{i \in S_k} \boldsymbol{\lambda}_s^i$ for structure s. Since we only have a few samples per class, we estimate a global diagonal covariance matrix Σ_s across all scans for each structure. Weighting distances by the variance helps to prevent the domination by higher eigenvalues that exhibit higher variation. The subject identity with the highest probability is assigned to the scan

$$k^* = \arg\max_k p(\mathcal{C}_k|\Lambda). \tag{2}$$

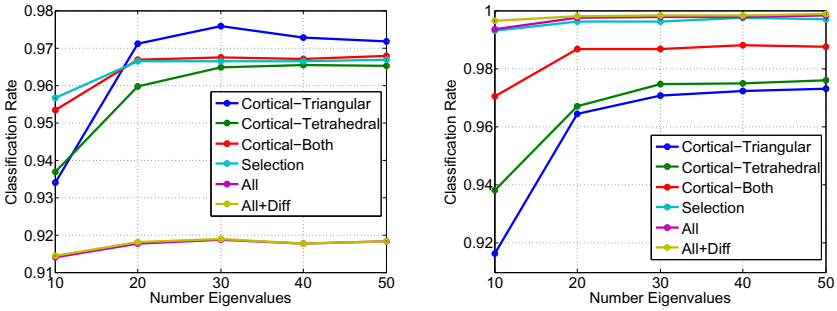


Fig. 3. Classification results for product classifier (left) and voting classifier (right) under variable number of eigenvalues and feature sets

The posterior probability of this classifier is the product of the posterior probabilities across all structures, cf. Eq. (1), which may be problematic for structures with low discriminative power. Many subcortical structures do not carry much distinctive shape information and can therefore negatively influence the overall probability. We therefore propose a second classifier that is specifically adapted to working with structures that are not very discriminative. Increased robustness is achieved by voting for each structure independently

$$k_s^* = \arg\max_k p(\boldsymbol{\lambda}_s|\mathcal{C}_k), \quad \forall s \in \{1,\ldots,\eta\}, \tag{3}$$

with the final vote set to the mode of the vote distribution.

4 Results

We perform experiments on data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [7]. We work with over 3000 scans from almost 700 subjects, where each subject has between three and six longitudinal scans. Each T1-weighted image from the dataset is processed independently with FreeSurfer. We calculate 36 shape descriptors for subcortical structures and 8 descriptors for cortical structures (left/right, white/gray matter, 2D/3D). Additionally, we calculate the lateral differences of shapeDNA between left and right cortical structures to quantify asymmetries, resulting in 4 additional descriptors.

We perform leave-one-out experiments by removing one scan from the dataset and by aiming to recover the correct identity. Fig. 3 reports the classification results for the product classifier in Eq.(2) and the structure-specific voting in Eq.(3). We report classification results as a function of the number of eigenvalues used to represent the shape. Additionally, we vary the set of brain structures in *BrainPrint*: cortical structures with triangular meshes (4), cortical structures with tetrahedral meshes (4), cortical structures for both mesh types (8), a selection of structures with the highest individual performances (15),

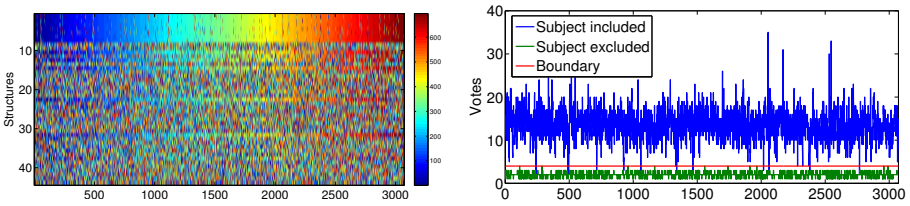


Fig. 4. Left: Subject (color) voted for by each structure (row) for each scan (column). Cortical structures in first 8 rows, subcortical features below. Optimal feature response would show a color gradient from blue to red, since scans are sorted by subject. Right: Number of votes for the winning subject identity when the correct subject is included (blue) in the database and when it is excluded (green). Decision boundary at 4 votes (red) yields a 0.49% false negative rate.

all structures (44), and all structures with the lateral differences of cortical structures (48). The number of structures is shown in parentheses. The results demonstrate a clear difference between the two classifiers. The product classifier achieves the best performance when working with cortical triangular meshes. Adding more features, especially when working with all features, dramatically reduces the classification results. We observe an opposite behavior when working with the structure-specific voting. Subcortical structures alone yield the worst performance in this case. The combination of 3D solid and 2D boundary descriptors leads to a clear improvement. A further improvement is gained by adding subcortical structures.

To further study this behavior, we examine the candidate subject that each structure votes for in Fig. 4. Each column corresponds to one scan and each row to one structure. The color indicates the subject number. Scans were sorted by subject; a perfect feature should show a color gradient from blue to red. The first 8 rows correspond to cortical structures, which exhibit the best performance. The remaining 36 rows show subcortical structures that perform worse than cortical structures and vary in their discriminative power. This explains the poor performance of the product classifier for the whole feature set, as weak features can obscure good features. In contrast, weak features do not degrade the performance of the voting classifier as long as weak features show no bias for a specific subject. The best performance of over 99.8% is achieved for 50 eigenvalues on all features with the additional difference features. For comparison, the classification rate for the mean LGI on both hemispheres is 1.0% for the product and 3.9% for the voting classifier. The classification rate for the volume, calculated from all cortical and subcortical structures, is 0.03% for the product and 0.6% for the voting classifier, confirming results from Fig. 1.

Fig. 5 shows the two scans for which *BrainPrint* does not correctly identify the subject identity. These subjects show strong atrophy and imaging artifacts, resulting in pronounced segmentation errors. Manual correction in FreeSurfer or reacquisition to avoid motion artifacts may therefore improve the above results.

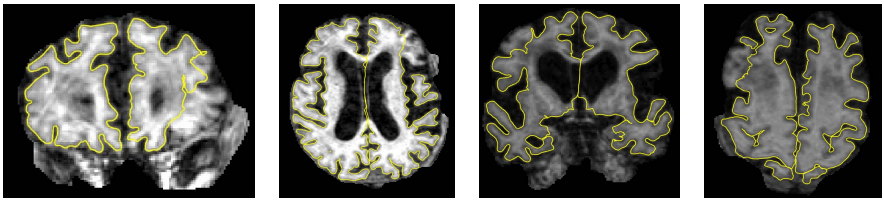


Fig. 5. Coronal and axial slices from two misclassified scans. White matter segmentation is shown in yellow.

As an additional experiment, we evaluate the possibility to determine whether a subject is not contained in the database. We study the number of votes the winning subject receives in Fig. 4, once when the subject of the scan is included in the database and once when the subject is excluded. If the subject in the current scan exists in the database, the scan receives about 15 votes for the winning subject class. If the subject is not contained in the database, the number of votes for the winner does not surpass 4. Setting 4 votes as our decision boundary results in only a 0.49% error (false negative) of concluding incorrectly that a subject is not in the database. The false positive rate is zero.

5 Discussion and Conclusions

The high classification accuracy of *BrainPrint* suggests that brain structures are unique to individuals and can be used for identification. Since our study only includes data on subjects followed over a period of up to 36 months, we cannot currently assess how the accuracy of *BrainPrint* changes across the entire lifespan of a subject. Unfortunately, such data sets are not yet available. However, since subjects with Alzheimer's disease in our dataset demonstrate pronounced neurodegeneration in a relatively short time, we are optimistic that *BrainPrint* will remain robust for comparison across longer time periods.

The identification accuracy may raise concerns about privacy issues when publicly distributing de-faced or skull-stripped brain scans together with diagnosis and other sensitive information. Yet, we currently do not think that *BrainPrint* interferes with anonymization because at least a second scan with knowledge of the identity needs to be available to connect to the private information. In terms of its practical applications, we see *BrainPrint* as an aid when handling large datasets. Identifying similar images in an efficient way can provide the launchpad for a more detailed follow-up analysis, *e.g.*, calculation or prediction of localized growth and shrinkage patterns. Since most of our retrieval errors are related to incorrect segmentations, our approach could also be used as an automatic quality control. Furthermore, *BrainPrint* can help identify anonymization errors (mismatch of subject identity), which are difficult to detect and can impede longitudinal studies. Finally, the presented framework of image understanding and compact characterization is relevant for handling large datasets in other fields and not limited to neuroscience.

Acknowledgements. This work was supported in part by the Humboldt foundation, the Martinos Center for Biomedical Imaging (P41-RR014075, P41-EB015896), the National Alliance for Medical Image Computing (U54-EB005149) and the NeuroImaging Analysis Center (P41-EB015902). We thank Anna Rieckmann for revising the manuscript and the Alzheimer's Disease Neuroimaging Initiative (ADNI) for image data.

References

1. Bates, J., Pafundi, D., Kanel, P., Liu, X., Mio, W.: Spectral signatures of point clouds and applications to detection of alzheimer's disease through neuroimaging. In: IEEE International Symposium on Biomedical Imaging, pp. 1851–1854 (2011)
2. Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M.: Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33(3), 341–355 (2002)
3. Gerber, S., Tasdizen, T., Fletcher, P.T., Joshi, S., Whitaker, R.: Manifold modeling for brain population analysis. Medical Image Analysis 14(5), 643–653 (2010)
4. Geuzaine, C., Remacle, J.F.: Gmsh: A 3-d finite element mesh generator with built-in pre-and post-processing facilities. International Journal for Numerical Methods in Engineering 79(11), 1309–1331 (2009)
5. Hamm, J., Ye, D.H., Verma, R., Davatzikos, C.: Gram: A framework for geodesic registration on anatomical manifolds. Med. Image Analysis 14(5), 633–642 (2010)
6. Lian, Z., Godil, A., Bustos, B., Daoudi, M., Hermans, J., Kawamura, S., Kurita, Y., Lavoué, G., Van Nguyen, H., Ohbuchi, R., et al.: A comparison of methods for non-rigid 3D shape retrieval. Pattern Recognition 46, 449–461 (2012)
7. Mueller, S.G., Weiner, M.W., Thal, L.J., et al.: The alzheimer's disease neuroimaging initiative. Neuroimaging Clinics of North America 15(4), 869–877 (2005)
8. Niethammer, M., Reuter, M., Wolter, F.E., Bouix, S., Peinecke, N., Koo, M.S., Shenton, M.: Global medical shape analysis using the Laplace-Beltrami spectrum. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part I. LNCS, vol. 4791, pp. 850–857. Springer, Heidelberg (2007)
9. Reuter, M., Niethammer, M., Wolter, F.E., Bouix, S., Shenton, M.: Global medical shape analysis using the volumetric Laplace spectrum. In: International Conference on Cyberworlds, NASA-GEM Workshop, pp. 417–426 (2007)
10. Reuter, M., Wolter, F.E., Peinecke, N.: Laplace-Beltrami spectra as "Shape-DNA" of surfaces and solids. Computer-Aided Design 38(4), 342–366 (2006)
11. Reuter, M., Wolter, F.E., Shenton, M., Niethammer, M.: Laplace-Beltrami eigenvalues and topological features of eigenfunctions for statistical shape analysis. Computer-Aided Design 41(10), 739–755 (2009)
12. Schaer, M., Cuadra, M.B., Tamarit, L., Lazeyras, F., Eliez, S., Thiran, J.: A surface-based approach to quantify local cortical gyrification. IEEE Transactions on Medical Imaging 27(2), 161–170 (2008)

Diffeomorphic Shape Trajectories for Improved Longitudinal Segmentation and Statistics

Prasanna Muralidharan[1], James Fishbaugh[1], Hans J. Johnson[2], Stanley Durrleman[3], Jane S. Paulsen[2], Guido Gerig[1], and P. Thomas Fletcher[1]

[1] School of Computing & SCI Institute, University of Utah, Salt Lake City, UT, USA
[2] Department of Psychiatry, Carver College of Medicine, University of Iowa, Iowa City, USA
[3] Inria Paris-Rocquencourt, Inserm U1127, CNRS UMR 7225, Sorbonne Universités, UPMC Univ Paris 06 UMR S 1127, ICM, Paris, France*

Abstract. Longitudinal imaging studies involve tracking changes in individuals by repeated image acquisition over time. The goal of these studies is to quantify biological shape variability within and across individuals, and also to distinguish between normal and disease populations. However, data variability is influenced by outside sources such as image acquisition, image calibration, human expert judgment, and limited robustness of segmentation and registration algorithms. In this paper, we propose a two-stage method for the statistical analysis of longitudinal shape. In the first stage, we estimate diffeomorphic shape trajectories for each individual that minimize inconsistencies in segmented shapes across time. This is followed by a longitudinal mixed-effects statistical model in the second stage for testing differences in shape trajectories between groups. We apply our method to a longitudinal database from PREDICT-HD and demonstrate our approach reduces unwanted variability for both shape and derived measures, such as volume. This leads to greater statistical power to distinguish differences in shape trajectory between healthy subjects and subjects with a genetic biomarker for Huntington's disease (HD).

1 Introduction

Statistical shape modeling and analysis is of critical importance for better understanding of longitudinal imaging and shape data, especially in the context of dynamic processes like aging and disease progression. To model evolution of shape, many regression approaches for cross-sectional data have been proposed [1–4]. However, regression has limitations when applied to longitudinal analysis, since each individual could start at a different point and evolve in a different manner. Longitudinal studies therefore entail development of subject-specific spatiotemporal models, and also a way to compare these models across different subjects [5–8].

* This research was supported by NIH Grants U01 NS082086, NS40068, NS050568 (PREDICT-HD), U54 EB005149 (NA-MIC), S10 RR023392 (NCCR Shared Instrumentation Grant), and NSF CAREER Grant 1054057. Also supported by NIH (NINDS; 5RO1NS040068, 5RO1NS054893) and the CHDI Foundation to Jane S. Paulsen. We thank the PREDICT-HD sites, the study participants, the National Research Roster for HD Patients and Families, the Huntington Disease Society of America and the Huntington Study Group.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 49–56, 2014.

Longitudinal image data has several sources of variability. First, there is inherent biological variability, both within a subject changing over time and also between subjects in a population. The goal of longitudinal analysis is to quantify this variability and make inferences about changes over time in a population. However, longitudinal imaging data also include unwanted sources of variability, such as noise in image acquisition, segmentation and registration errors, and human expert judgment, among others. These extraneous errors tend to dampen statistical power, especially when trying to distinguish between trajectories of two different populations, e.g., healthy and diseased.

In this paper, we propose a framework that first seeks to reduce this extraneous variability, thus improving consistency of longitudinal segmentations in the first stage. We follow the procedure of [9] by estimating diffeomorphic geodesic trajectories of shape evolution for each individual. The estimated trajectories are smooth, resulting in temporally consistent and more biologically plausible shape evolution. We then employ a mixed effects model for shapes [10, 7] to conduct longitudinal statistical shape analysis on the consistent shape trajectories. We demonstrate the benefit of our two stage approach by a comparison of longitudinal mixed-effects analysis on cortical volumes obtained from raw observed data against consistent measurements obtained from personalized spatiotemporal shape models. We also show our method reduces unwanted variability for both shape and derived measures, such as volume. This leads to greater statistical power to distinguish shape evolution between healthy subjects and subjects with a genetic biomarker for Huntington's disease (HD).

2 Methodology

We present here methodology for the statistical analysis of longitudinal shape complexes. This is based on spatiotemporal modeling of diffeomorphic shape trajectories (Section 2.1) to produce temporally consistent shape sequences. Estimated model trajectories represent more biologically plausible and smooth shape changes associated with anatomical evolution in time. Statistical measures and group hypothesis testing is then conducted on both scalar measurements extracted from shape as well as the shape complexes themselves. For measuring individual and group shape differences, we estimate a multivariate mixed-effects model (Section 2.2) for shapes, designed to take advantage of longitudinal shape data.

2.1 Spatiotemporal Modeling for Consistency in Longitudinal Segmentation

Anatomical change over time associated with neurodevelopment or aging is assumed to be a smooth process. That is, the trajectory of a particle on an anatomical surface should be differentiable, with no instantaneous change of direction. The presence of a disorder such as Huntington's disease (HD) would not invalidate the smoothness assumption. Rather, the neurodegeneration process associated with HD has been observed as a temporally smooth process [11]. However, our anatomical measurements (medical images and extracted anatomical shapes) are often not representative of samples from a smooth process, due to the natural variability attributed to image acquisition, subject positioning, segmentation, etc. Without temporal consistency in our measurements,

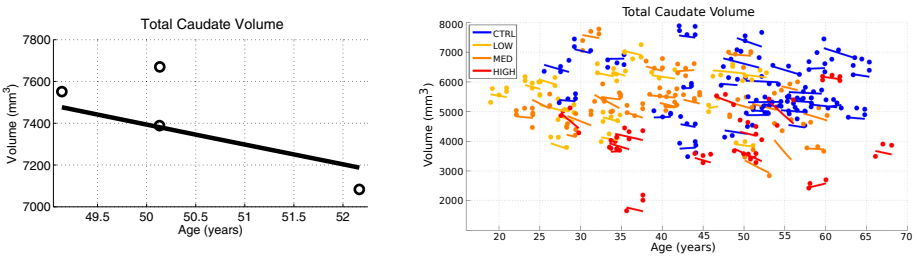


Fig. 1. Left: For one subject, volume of observed caudates (open circles) and temporally consistent continuous volume extracted from diffeomorphic shape model (solid line). The difference in caudate volume extracted from scans obtained on the same day highlights the need for consistent segmentation. Right: Observed volume and volume extracted from diffeomorphic shape models for all 65 subjects. While the volume of the discrete shape observations show considerable variation, volume extracted from personalized models are continuous and temporally consistent.

it becomes difficult to distinguish between anatomical change associated with disease from changes due to noise.

One emerging model of smooth anatomical change is to consider continuous transformations of the ambient space by differentiable and invertible deformations. We model anatomical trajectories by a geodesic flow of diffeomorphisms that continuously deforms a given anatomical configuration $\mathbf{X}_0$ over time to closely match a set of observed anatomical shapes $\mathbf{O}_{t_i}$ [9]. The initial anatomical configuration (baseline shape), as well the flow of diffeomorphisms ϕ_t, are estimated by minimizing the criterion

$$E(\mathbf{X}_0, \phi_t) = \sum_i D(\phi_{t_i}(\mathbf{X}_0) - \mathbf{O}_{t_i}) + \mathrm{Reg}(\phi_t),$$

where D represents a distance metric on shapes and $\mathrm{Reg}(\phi_t)$ is a measure of the regularity of the geodesic flow of diffeomorphisms ϕ_t. For choice of D, we favor the metric on currents, which is robust to topological differences and allows for comparison between shapes without the need for point correspondence. Also, being in an infinite-dimensional space of diffeomorphisms, geodesic trajectories have the flexibility to capture complex deformations.

The continuous geodesic flow of diffemorphisms ϕ_t is applied to the estimated anatomical configuration to produce a continuous and temporally consistent sequence of shapes. The improved temporal consistency is illustrated on the left side of Fig. 1 by comparing the volume of observed caudates with the volume extracted *continuously* from the spatiotemporal model of caudate shape. Also note that we can now obtain shapes or measurements extracted from shapes at any time point of interest, not just those corresponding to observations.

2.2 Mixed Effects Model for Shapes

We now have a diffeomorphic flow of anatomical shapes for each individual, from which we obtain shapes at time points corresponding to actual observations. These

estimated shapes no longer represent independent (and potentially noisy) measurements, but instead take into account correlation between repeated scans of the same individual.

Statistical interpretation of longitudinal shape data is extremely useful in ascertaining differences in repeated image scans of an individual and also between individuals within and across populations. A compact statistical representation of shape was proposed by [10], wherein the surface of a shape is represented by a collection of points, also referred to as a particle system. Particle positions are optimized to be in correspondence across an ensemble of shape configurations. A faithful shape representation is achieved by minimizing a cost function, that balances a low residual error of model to data, also seeking configurations of uniformly-distributed correspondence positions on shape surfaces.

To analyze longitudinal data, [7] generalize the methods in [10] to incorporate a linear mixed-effects model in the optimization framework. Let Y_i be the longitudinal response variable for the ith individual (a shape configuration), and X_i denote the explanatory variable, typically time. The mixed-effects model for longitudinal correspondences is given as

$$Y_i = X_i(\alpha + b_i) + \epsilon_i,$$

where α are the fixed-effects parameters (group intercept, group slope), while b_i are random-effects parameters with ϵ_i being the error in correspondences for the ith individual. For details on model parameter estimation, see [7].

Hypothesis Testing. In order to test the statistical significance of group-parameter differences between two groups of longitudinal data, [7] also outline a statistical hypothesis permutation test based on the Hotelling's T^2 statistic.

Given two groups of data, $\{p_1, \ldots, p_m\}$ and $\{q_1, \ldots, q_n\}$, with sample means $\bar{p}, \bar{q}$, recall that Hotelling's T^2 statistic is a test statistic to test for significant differences between sample means, relative to the pooled sample covariance W:

$$W = \frac{\sum_i (p_i - \bar{p})(p_i - \bar{p})^T + \sum_i (q_i - \bar{q})(q_i - \bar{q})^T}{m + n - 2}.$$

The T^2 statistic can be thought of as a squared Mahalanobis distance between the means, using the pooled covariance W. The sample T^2 statistic is given by

$$t^2 = \frac{mn}{m+n}(\bar{p} - \bar{q})^T W^{-1} (\bar{p} - \bar{q}).$$

The permutation test procedure is as follows: (1) compute the t^2 statistic, (2) randomly permute (swap) data points between the p and q groups, computing a t_k^2 statistic for the permuted groups, (3) repeat step 2 for $k = 1, \ldots, P$, (4) compute the p-value: $p = B/(P+1)$, where B is the number of $t_k^2 < t^2$. The final p-value can be interpreted as the probability of finding a larger group difference by random chance under the null hypothesis (that there is no difference between the means). The underlying assumption of any permutation test is that the data should be exchangeable under the null distribution. Our null hypothesis is that the groups (e.g., healthy and diseased) are from the

same distribution. We permute individuals (keeping their timepoints all intact), which under this null assumption is exchangeable.

To test for differences in anatomical trajectories between a healthy and disease group, also note that it is important to distinguish if the shape differences are present at baseline (intercept) or if they develop over time (slope). To make this distinction, we also separate the above Hotelling's T^2 test into these two components.

3 Experimental Validation

We study subcortical change associated with Huntington's disease (HD), leveraging the longitudinal study PREDICT-HD. The longitudinal database consists of 65 female subjects: 23 controls (CTRL), 14 (LOW), 15 (MED), and 13 (HIGH). The LOW / MED / HIGH categories represent probability of onset of manifesting signs of HD. All subjects have had at least 3 MR images acquired approximately one year apart, with many subjects undergoing multiple scans per visit. Six subcortical pairs (caudate, putamen, hippocampus, thalamus, acumben, and pallidus) were segmented from each image (Fig. 2) and manually verified and cleaned [12].

The quality of each segmentation varies considerably for each time point, even when scans are obtained on the same day from the same scanner, as individual single-subject segmentation is prone to errors related to variability of imaging, image calibration, human expert judgment, and limited robustness of segmentation algorithms. While the segmentation quality is not easily assessed by viewing the 3D anatomical surfaces, the temporal inconsistency becomes clear by investigating volume extracted from the shapes. The right side of Fig. 1 shows the variability in segmentation, illustrated by the temporal inconsistency of observed caudate volume, motivating the need for temporally consistent segmentations which properly account for correlated longitudinal data.

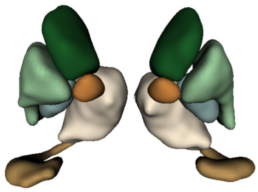


Fig. 2. Example of six subcortical pairs extracted for each subject and timepoint

Personalized Spatiotemporal Models of Subcortical Change. Continuous models of shape trajectory are estimated for each subject using the methodology outlined in 2.1, resulting in personalized and temporally consistent anatomical evolution. Model estimation does not require point correspondence, facilitating the inclusion of all subcortical shapes simultaneously without imposing any topological constraints. Each subject's personalized model allows us to generate shapes at any instant in time, from which desired shape properties, such as volume, can be extracted. We can therefore obtain a continuous evolution of volume for all subcortical structures without any explicit modeling of volume. Fig. 1 shows caudate volume extracted from each subject's continuous shape model, demonstrating the flexibility of the shape model to capture both linear and non-linear volume trends with no prior assumption or constraint on linearity. Though we only display caudate volume here, recall that each model is estimated by leveraging all shape data simultaneously (Fig. 2), which respects shape boundaries and locations, incorporating important geometric relationships between shapes.

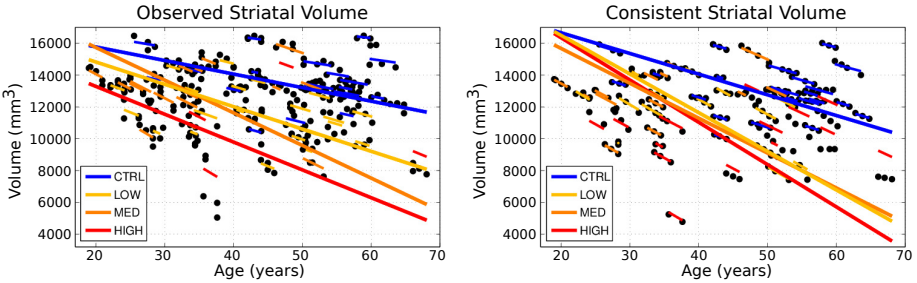


Fig. 3. Longitudinal mixed-effects analysis of striatal volumes obtained from observed shapes (Left) and temporally consistent shapes (Right). Volume data are shown as filled black circles with corresponding individual trends. Note the improvement of the model fit in the consistent striatal volume over the observed striatal volume, which results in lower standard error of estimated mixed-effects parameters. (See Table 1)

Longitudinal Analysis of Striatal Volume. Here we conduct a univariate analysis of volume extracted from shape, as striatal volume loss has been shown to be associated with the progression of HD [11]. We aim to evaluate the benefit of spatiotemporal shape modeling, by comparing striatal volume extracted from the temporally consistent shapes with volume extracted from the raw shape observations. Figure 3 shows the results of linear mixed-effects analysis on striatal volumes for observed (left) and temporally consistent shapes (right), testing for the interaction between age and group membership. The estimated fixed-effects parameters for the temporally consistent (smoothed) category were found to be significant, as shown in Table 1. This demonstrates the benefit of spatiotemporal shape modeling, as striatal volumes extracted from the temporally consistent shapes provide better separation between the control and LOW groups, and also between the control and HIGH groups.

Another benefit of spatiotemporal shape modeling is seen in the standard error of estimated parameters (Table 1). The standard error is consistently lower for temporally consistent shapes, which implies a reduction in unwanted variability present in the original segmentations. Further note in Fig. 3, the mixed-effects model fits the temporally consistent data better than the observed striatal volume. We also performed separate longitudinal mixed-effects analysis on the caudate and the putamen, and found a similar story in both cases.

Table 1. Comparison of the standard error and significance values of fixed-effects parameters of longitudinal volumes obtained from observed and temporally consistent shapes

Parameter	Std. error (obs.)	Std. error (smoothed)	p-value (obs.)	p-value (smoothed)
Fixed-effects (slope)	26.23	14.10	**0.002**	$<$ **0.001**
Slope (high)	66.56	23.73	0.182	$<$ **0.001**
Slope (med)	36.43	22.14	**0.003**	$<$ **0.001**
Slope (low)	38.73	26.60	0.143	$<$ **0.001**

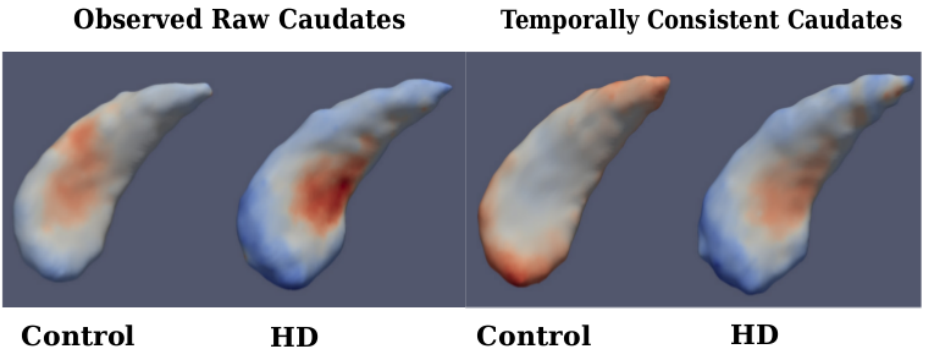


Fig. 4. Left: Fixed-effects parameters for observed caudate shapes (Far Left-Control, Mid Left-HD), Right: Fixed-effects parameters for temporally consistent caudate shapes (Mid Right-Control, Far Right-HD); Fixed effects slope: Blue-Red indicates Local Contraction - Expansion

Table 2. *p-values:* Hypothesis test for differences in shape change ("slope"), between controls and HD groups, for observed caudates (Left) and temporally consistent caudate shapes (right)

Structure	Observed	Temporally consistent
Left caudate	0.15	**0.005**
Right caudate	0.23	0.06

Longitudinal Analysis of Striatal Shape. We next perform a multivariate Hotelling's T^2 hypothesis test (Section 2.2) of the baseline shape (intercept) and trend (slope) between controls and the combined HD groups. We compare the results for analyses using the original observed segmentations versus those obtained from spatiotemporal modeling as described in (Section 2.1). We represent these shapes in the particle optimization framework to estimate longitudinal fixed and random effects. Note that we do not normalize for size in these experiments, which means that we test for differences between control and combined HD groups based on both shape and size.

Figure 4 shows the estimated fixed-effects parameters for both groups, i.e., the baseline (intercept) shape with trajectory (slope) displayed as a color map. When comparing baseline shapes, we don't find significant difference between controls and HD in either analysis. This is expected, as the onset of degeneration in HD is expected at a later age. But when comparing shape trends, we find significant differences between controls and HD for the temporally consistent shapes, but not in the case of raw shape observations.

Table 2 provides the p-values from the statistical hypothesis test between the control and combined HD groups. In both the left and right caudate, the temporally consistent shapes result in lower p-values. Specifically, the left caudate is statistically significant at the 5% level. Similar to the volume analysis, this demonstrates that temporally consistent shape trajectories result in greater ability to distinguish differences in longitudinal trends between controls and HD groups.

4 Conclusion

Diffeomorphic trajectories are good at capturing smooth anatomical shape changes, while the particle optimization framework excels at finding compact statistical shape representations with increased statistical power. The novelty of our work is to leverage the strengths of both approaches, to provide an integrated solution, characterized by improved statistical performance in the analysis of both scalar and shape trajectory data derived from noisy segmentations. We demonstrate the advantages of our method through improved statistics on temporally consistent shape and volume measures in the analysis of the PREDICT-HD dataset.

References

1. Davis, B., Fletcher, P.T., Bullitt, E., Joshi, S.: Population shape regression from random design data. In: Proceedings of IEEE International Conference on Computer Vision (2007)
2. Niethammer, M., Huang, Y., Vialard, F.X.: Geodesic regression for image time-series. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part II. LNCS, vol. 6892, pp. 655–662. Springer, Heidelberg (2011)
3. Fletcher, P.T.: Geodesic regression and the theory of least squares on riemannian manifolds. Int. J. Comput. Vision 105(2), 171–185 (2013)
4. Hinkle, J., Muralidharan, P., Fletcher, P.T., Joshi, S.: Polynomial regression on Riemannian manifolds. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 1–14. Springer, Heidelberg (2012)
5. Qiu, A., Albert, M., Younes, L., Miller, M.: Time sequence diffeomorphic metric mapping and parallel transport track time-dependent shape changes. NeuroImage 45, S51–S60 (2009)
6. Muralidharan, P., Fletcher, P.T.: Sasaki metrics for the analysis of longitudinal data on manifolds. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
7. Datar, M., Muralidharan, P., Kumar, A., Gouttard, S., Piven, J., Gerig, G., Whitaker, R., Fletcher, P.T.: Mixed-effects shape models for estimating longitudinal changes in anatomy. In: Durrleman, S., Fletcher, T., Gerig, G., Niethammer, M. (eds.) STIA 2012. LNCS, vol. 7570, pp. 76–87. Springer, Heidelberg (2012)
8. Durrleman, S., Pennec, X., Trouvé, A., Braga, J., Gerig, G., Ayache, N.: Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data. Int. J. Comput. Vision 103(1), 22–59 (2013)
9. Fishbaugh, J., Prastawa, M., Gerig, G., Durrleman, S.: Geodesic Shape Regression in the Framework of Currents. In: Gee, J.C., Joshi, S., Pohl, K.M., Wells, W.M., Zöllei, L. (eds.) IPMI 2013. LNCS, vol. 7917, pp. 718–729. Springer, Heidelberg (2013)
10. Cates, J.E., Fletcher, P.T., Styner, M.A., Shenton, M.E., Whitaker, R.T.: Shape modeling and analysis with entropy-based particle systems. In: Karssemeijer, N., Lelieveldt, B. (eds.) IPMI 2007. LNCS, vol. 4584, pp. 333–345. Springer, Heidelberg (2007)
11. Aylward, E., Mills, J., Liu, D., Nopoulos, P., Ross, C.A., Pierson, R., Paulsen, J.S.: Association between Age and Striatal Volume Stratified by CAG Repeat Length in Prodromal Huntington Disease. PLoS Curr. 3, RRN1235 (2011)
12. Kim, E.Y., Johnson, H.J.: Robust multi-site mr data processing: Iterative optimization of bias correction, tissue classification, and registration. Frontiers in Neuroinformatics 7(29) (2013)

Simulating Neurodegeneration through Longitudinal Population Analysis of Structural and Diffusion Weighted MRI Data

Marc Modat[1,2], Ivor J.A. Simpson[1,2], Manuel Jorge Cardoso[1,2], David M. Cash[2,1], Nicolas Toussaint[1,2], Nick C. Fox[2], and Sébastien Ourselin[1,2]

[1] Translational Imaging Group, Centre for Medical Imaging Computing, Department of Medical Physics and Bioengineering, University College London, UK
[2] Dementia Research Centre, Institute of Neurology, University College London, Queen Square, WC1N 3BG, UK

Abstract. Neuroimaging biomarkers play a prominent role for disease diagnosis or tracking neurodegenerative processes. Multiple methods have been proposed by the community to extract robust disease specific markers from various imaging modalities. Evaluating the accuracy and robustness of developed methods is difficult due to the lack of a biologically realistic ground truth.

We propose a proof-of-concept method for a patient- and disease-specific brain neurodegeneration simulator. The proposed scheme, based on longitudinal multi-modal data, has been applied to a population of normal controls and patients diagnosed with Alzheimer's disease or frontotemporal dementia. We simulated follow-up images from baseline scans and compared them to real repeat images. Additionally, simulated maps of volume change are generated, which can be compared to maps estimated from real longitudinal data. The results indicate that the proposed simulator reproduces realistic patient-specific patterns of longitudinal brain change for the given populations.

1 Introduction

Imaging biomarkers have a strong potential to aid disease diagnosis, tracking changes over time or to evaluate new treatments. Within neurodegenerative disorders such as Alzheimer's disease (AD) there are numerous robust techniques routinely used to quantify atrophy [1, 2] or to characterise disease specific patterns of brain changes [3]. There is however no method to assess the accuracy and robustness of these techniques due to the lack of ground truth. Numerous proposals have been made to simulate the longitudinal changes observed in structural MRI. Davatzikos *et al.* [4] simulated brain atrophy by introducing a 30% uniform volume change in two specific gyri. Camara *et al.* [5] took advantage of regional rates of atrophy reported in the literature to simulate the expected change to a baseline T1 weighted (T1w) MRI scan that would be observed at follow-up for an AD subject. In this work, a finite element representation of a template brain was registered to a new subject scan in order to propagate the

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 57–64, 2014.

tetrahedral mesh. Regional volume changes are applied to the propagated mesh to simulate through linear elasticity a transformation which is then applied to the T1w MR image. Sharma *et al.* [6] propose to generate a learning database from simulated images with various amounts of atrophy and noise, from which they then assessed a ground truth atrophy.

All of these simulation methods [4–6] assume a common atrophy rate across all subjects. Camara *et al.*'s method is specific to the simulation of an AD pattern of atrophy, whereas the other methods do not relate specifically to a given disease and only describe volume changes within a sub-region of the brain (specific gyri for Davatzikos *et al.* and hippocampi for Sharma *et al.*).

We present a proof-of-concept method for a patient- and disease-specific neurodegeneration simulator. Pattern of longitudinal changes are learned from a template database that consists of MRI scans from normal controls, subjects diagnosed with frontotemporal dementia or sporadic AD patients. For each subject in the database, we have access to baseline and 1 year follow-up scans. Using multi-modal[1] image based registration, we first extract velocity fields that characterise the longitudinal changes occurring over a year for each subject. The extracted information is then combined within the space of a new subject to simulate a patient- and disease-specific flow of deformation. Here, we hypothesise that brains with similar morphology evolve similarly. Image registration is known to be an ill-posed problem and the obtained result are hence heavily dependant of algorithmic choices such as transformation model and regularisation. In order to reduce this bias and the variance of all possible solutions, we take advantage of a registration scheme combining structural and diffusion MRI as in Avants *et al.* [7] or Studholme [8].

2 Method

Within the proposed neurodegeneration simulator scheme, we extract information from a template database containing MRI acquisitions from 40 normal controls (NC), 17 patients diagnosed with frontotemporal dementia (FTD) and 17 patients diagnosed with Alzheimer's disease (AD). The overall pipeline of the proposed framework is presented in Figure 1.

2.1 Data and Pre-processing

Baseline and 1 year follow-up scans are available for each subject. Each time point consists of a pair of images: a structural T1w MRI and a DWI scan both acquired on a Siemens Tim Trio 3 Tesla scanner using a 32-channel head coil. The DWI scans were acquired with 64 directions (b-value=1000) repeated twice and 9 b0 images were acquired. On the diffusion images, we performed motion and eddy current correction as well as EPI correction using a scheme combining field maps and non-linear registration. Both 64 direction acquisitions were used

[1] We should refer to multi-pulse sequences but use multi-modal for simplicity.

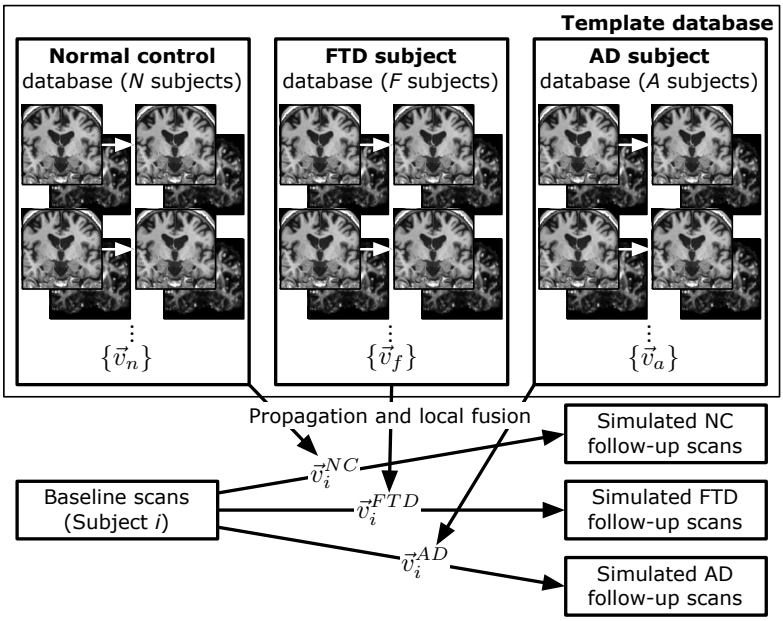


Fig. 1. Framework of the overall pipeline. Starting from a template database containing longitudinal multi-modal MRI scans from normal controls (NC), subjects with frontotemporal dementia (FTD) and Alzheimer's disease (AD), intra-subject longitudinal flow of deformations are extracted through non-linear registration. The flows are then propagated and fused to create new flows of deformation simulating, for a given subject, change over time as if the subject was NC, FTD or AD.

concurrently to fit a tensor per voxel. The geometry distortion due to gradient non-linearity on the structural image was corrected, as well as the intensity nonuniformity [9]. The average b0 image was rigidly registered to its corresponding structural T1w image and the obtained transformation was used to resample the tensor image into the T1w space using a log-Euclidean interpolation scheme.

2.2 Multi-channel Intra-subject Registration Using Multi-modal Information

As mentioned, non-linear registration lacks a unique optimal solution and usually relies on mathematical formulations for regularisation. In order to decrease the influence of such a formulation, we take advantage of multi-modal imaging to increase the amount of information available to drive the transformation. As in Avants *et al.* [7] and Studholme [8], we use structural and diffusion MRI images.

The proposed scheme for registration, implemented within the NiftyReg package [10], is based on a cubic b-spline parametrisation of a continuous stationary velocity field, $\boldsymbol{v}$, which yields a deformation field, $\boldsymbol{u}$, through exponentiation using a scaling-and-squaring approach [11]. We use a symmetric scheme where

forward and backward deformation fields, $\boldsymbol{u}$ and $\boldsymbol{u}^{-1}$, are concurrently optimised. This is achieved thanks to the following property: $\exp(\boldsymbol{v}) = \exp(-\boldsymbol{v})^{-1}$ where both velocity fields, $\boldsymbol{v}$ and $-\boldsymbol{v}$, are estimated concurrently from the parameters $\{\mu\}$, the cubic b-spline coefficients. Using the T1w and DTI images for both baseline, B, and follow-up image, F, the measure of similarity, $\mathcal{M}$, used to drive the registration is computed as:

$$\begin{aligned}\mathcal{M}(B, F; \mu) = \alpha \times \mathcal{M}_s(B^s, F^s(\boldsymbol{u}; \mu)) + \alpha \times \mathcal{M}_s(B^s(\boldsymbol{u}^{-1}; \mu), F^s) \\ + \beta \times \mathcal{M}_d(B^d, F^d(\boldsymbol{u}; \mu)) + \beta \times \mathcal{M}_d(B^d(\boldsymbol{u}^{-1}; \mu), F^d),\end{aligned} \tag{1}$$

where $\mathcal{M}_s(.)$ corresponds to the locally normalised cross correlation summed over all voxels [12] and $\mathcal{M}_d(.)$ corresponds to distance between the deviatoric tensors also summed over all voxels as in Zhang *et al.* [13]. B^s and F^s denote the structural T1w image from the baseline and follow-up images whereas B^d and F^d denote their DTI components. The values α and β are set empirically to 0.5 and 0.5. The overall cost function to be optimised by the registration framework consists of the measure of similarity $\mathcal{M}$ and the sum over all voxels of the bending energy of the velocity fields, $\boldsymbol{v}$ and $-\boldsymbol{v}$. The spacing between the control points was empirically set to 5 voxels and a pyramidal approach was three levels was used. For all subjects $n \in N$, $f \in F$ and $a \in A$ with N, F and A the number of NC, FTD and AD subjects respectively in the template database, we obtained sets of stationary velocity fields $\{\boldsymbol{v}_n\}$, $\{\boldsymbol{v}_f\}$ and $\{\boldsymbol{v}_a\}$ characterising the intra-subject longitudinal changes.

2.3 Simulated Flow of Longitudinal Changes

Given a new subject i and its T1w and DTI combined image B_i, we register all baseline images from the template database to B_i using the registration scheme described in section 2.2. We thus obtain three sets of deformation fields $\{\boldsymbol{u}_n^i\}$, $\{\boldsymbol{u}_f^i\}$ and $\{\boldsymbol{u}_a^i\}$. These deformations are used to propagate the previously obtained intra-subject longitudinal stationary velocity field into the space of image B_i. Similarly to the concept of weighted label fusion, which characterises the distance between the warped template image and the target image [14], we create a subject- and disease-specific flow of deformation using an local weighted average of multiple longitudinal flows. The distance $\mathcal{D}$ between a target image B_i and a template image B_j with $j \in [0, N + F + A[$ is computed as:

$$\mathcal{D}(B_i, B_j; \boldsymbol{u}_j^i) = \alpha \times \mathcal{D}_s(B_i^s, B_j^s(\boldsymbol{u}_j^i)) + \beta \times \mathcal{D}_d(B_i^d, B_j^d(\boldsymbol{u}_j^i)), \tag{2}$$

where B^s and B^d denote the structural and the diffusion component respectively of an image B. $\mathcal{D}_s$ and $\mathcal{D}_d$ correspond to the distances between the structural and the diffusion component of B_i and B_j. Within the current experiment, we set $\mathcal{D}_s = (2 - \mathcal{M}_s)$ and $\mathcal{D}_d = G * \mathcal{M}_d$, where G is a Gaussian kernel defined to match the kernel used to compute the locally normalised cross-correlation and $*$

is the convolution operator. Based on this distance the new flows of deformation for subject i are computed as:

$$\boldsymbol{v}_i^{grp} = \frac{\sum_{n \in \text{grp}} \left(\boldsymbol{u}_n^i \circ \boldsymbol{v}_n \right) \times e^{-\mathcal{D}(B^i, B^n; \boldsymbol{u}_n^i)/t}}{\sum_{n \in \text{grp}} e^{-\mathcal{D}(B^i, B^n; \boldsymbol{u}_n^i)/t}}, \tag{3}$$

where $\circ$ is the composition operator and t is the Gaussian kernel density, which is set to 0.5 for the current experiment. The group under consideration: NC, FTD or AD is denoted as grp. Note that the stationary velocity fields were here resampled using a gradient reorientation scheme.

Finally, simulated follow-up images F_i^{NC}, F_i^{FTD} and F_i^{AD} can be simulated from an input image B_i as $F_i^{grp} = \exp(\boldsymbol{v}_i^{grp}) \circ B_i$. Note that the transformation is parametrised with a flow field enabling interpolation between 0 and 1 year and extrapolation after 1 year. Figure 2 shows an example of simulated images from the baseline of a normal control baseline scan.

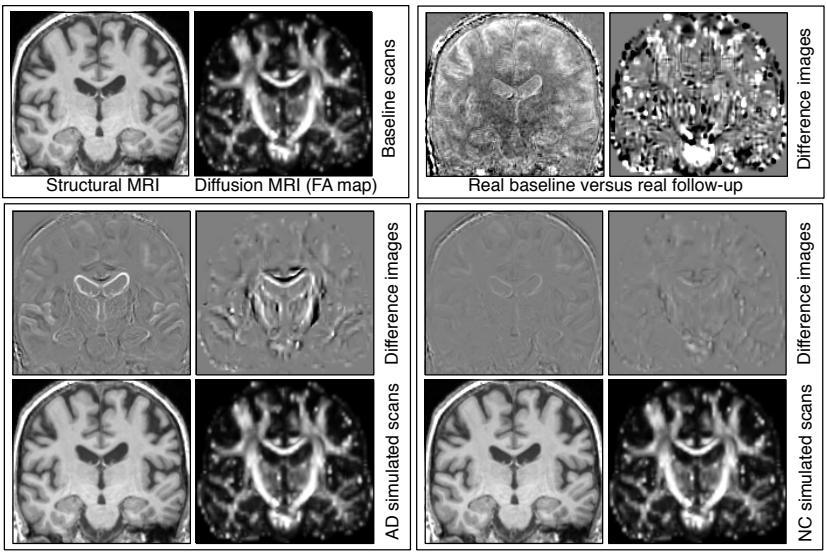


Fig. 2. Subject- and disease-specific longitudinal changes simulator result. Synthetic normal control and Alzheimer's disease structural and diffusion follow-up scans are simulated from the baseline T1w/DTI of a normal control subject. The simulated images are shown as well as the difference images between the simulated follow-up and the original scans. Note that for visualisation purpose the fractional anisotropy maps are used to visualise the diffusion information but DTI is used for the whole processing.

3 Evaluation

We performed a leave-one-out experiment where each subject from the aforementioned database was considered as a new subject and all the others composed the template database in a leave-one-out fashion. For each subject we simulated,

from its baseline scans, three T1w MR images and three DTIs, corresponding to the three groups in the template database, NC, FTD and AD. We were then able to evaluate, for each simulated scan its distance to the real one year follow-up image. The global normalised cross-correlation was used to quantify the distance between the real and the simulated structural follow-up images. This measure was selected to accomodate for potential global intensity scaling between the input images. The distance between the diffusion images was assessed using the mean of squared difference (MSD) between the fractional anisotropy (FA) maps derived from the real and the simulated DTIs. For both distance evaluations, skull-stripped images were used in order to only consider the voxels in the brain area. The result are presented in Figure 3. We also computed the distance,

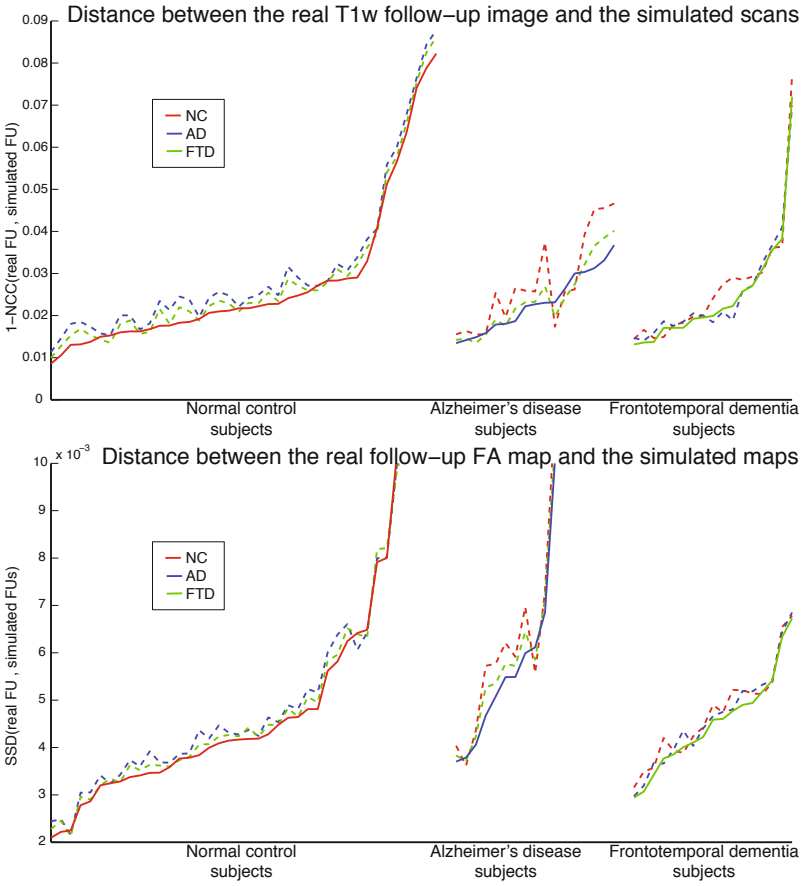


Fig. 3. Distance between the real and simulated follow-up structural (top) and diffusion (bottom) scans. For each baseline scan, we simulated three follow-up images corresponding to the three groups in the database: NC, FTD and AD. The distance between the real and the simulated scans are reported here. For visualisation purpose, distances have been sorted in an increasing order.

defined as the SSD, between the simulated logarithm of the Jacobian maps and the maps recovered by registering the real baseline and follow-up pairs of images.

It can be noticed that the differences between the different simulations are relatively small when compared to the actual distances between simulated and real data. This difference arises from other sources of variability such as image noise, artefacts or tensor fit errors, not related to geometrical differences. Moreover, the distances are computed over the full brain area whereas the simulated transformations impact largely on smaller regions such as the ventricles and the cortex in the T1w and the tract areas in the DTI. This is illustrated by the difference images between the real baseline and follow-up acquisitions in Figure 2. Nonetheless, 77.5%, 75% and 100% of the simulated NC are the closest to the real NC follow-up T1w/FA and estimated Jacobian maps. The equivalent numbers for the AD and FTD subjects are 64.7%/52.9%/88.2% and 58.8.7%/58.8%/88.2% respectively. This indicates that the proposed methodology captures disease specific patterns of longitudinal changes.

4 Discussion

We proposed a neurodegeneration simulator based on longitudinal and multimodal imaging data. We hypothesise that the use of multi-modal data decreases the variance of all possible registration results and increases the biological plausibility of the obtained transformations. However further assessment is needed to fully investigate the added value of using multi-modal registration rather than single modality registration. Using a local weighted scheme, we ensured subject specific simulation where the information is propagated only between similar morphologies. Due to the exploratory nature of the proposed work, further validation will be performed to assess, for example, the influence of the different algorithmic parameters, such as the weighting of the distance measure. The propagation of flow from one image space to another space was here performed using a gradient reorientation scheme. We will investigate the use of parallel transport techniques, such as pole ladder [15]. Finally, as previously mentioned, this work focused on the geometrical component of the longitudinal changes and not on the metamorphic process of pathology, which we will also further investigate.

Acknowledgment. M. Modat is supported by the UCL Leonard Wolfson Experimental Neurology Centre. D.M Cash is in part supported through a grant from Brain Research Trust. S. Ourselin receives funding from the EPSRC (EP/H046410/1, EP/J020990/1, EP/K005278), the MRC (MR/J01107X/1), the EU-FP7 project VPH-DARE@IT (FP7-ICT-2011-9-601055), the NIHR Biomedical Research Unit (Dementia) at UCL and the National Institute for Health Research University College London Hospitals Biomedical Research Centre (NIHR BRC UCLH/UCL High Impact Initiative). The Dementia Research Centre is supported by Alzheimer's Research UK, Brain Research Trust, and The Wolfson Foundation.

References

1. Smith, S.M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P.M., Federico, A., De Stefano, N.: Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. NeuroImage 17(1), 479–489 (2002)
2. Leung, K.K., Ridgway, G.R., Ourselin, S., Fox, N.C., Initiative, A.D.N.: Consistent multi-time-point brain atrophy estimation from the boundary shift integral. NeuroImage 59(4), 3995–4005 (2012)
3. Chung, M., Worsley, K.J., Paus, T., Cherif, C., Collins, D.L., Giedd, J.N., Rapoport, J.L., Evans, A.C.: A unified statistical approach to deformation-based morphometry. NeuroImage 14(3), 595–606 (2001)
4. Davatzikos, C., Genc, A., Xu, D., Resnick, S.M.: Voxel-based morphometry using the ravens maps: methods and validation using simulated longitudinal atrophy. NeuroImage 14(6), 1361–1369 (2001)
5. Camara, O., Schweiger, M., Scahill, R., Crum, W., Sneller, B., Schnabel, J., Ridgway, G., Cash, D., Hill, D.L.G., Fox, N.: Phenomenological model of diffuse global and regional atrophy using finite-element methods. IEEE Transactions on Medical Imaging 25(11), 1417–1430 (2006)
6. Sharma, S., Rousseau, F., Heitz, F., Rumbach, L., Armspach, J.P.: On the estimation and correction of bias in local atrophy estimations using example atrophy simulations. Computerized Medical Imaging and Graphics 37(7-8), 538–551 (2013)
7. Avants, B.B., Duda, J.T., Zhang, H., Gee, J.C.: Multivariate normalization with symmetric diffeomorphisms for multivariate studies. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part I. LNCS, vol. 4791, pp. 359–366. Springer, Heidelberg (2007)
8. Studholme, C.: Dense feature deformation morphometry: Incorporating DTI data into conventional MRI morphometry. Medical Image Analysis 12(6), 742–751 (2008)
9. Sled, J., Zijdenbos, A., Evans, A.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Transactions on Medical Imaging 17(1), 87–97 (1998)
10. Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S.: Fast free-form deformation using graphics processing units. Comput. Meth. Prog. Bio. 98(3), 278–284 (2010)
11. Arsigny, V., Commowick, O., Pennec, X., Ayache, N.: A log-Euclidean framework for statistics on diffeomorphisms. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 924–931. Springer, Heidelberg (2006)
12. Cachier, P., Bardinet, E., Dormont, D., Pennec, X., Ayache, N.: Iconic feature based nonrigid registration: the PASHA algorithm. Computer Vision and Image Understanding 89(2-3), 272–298 (2003)
13. Zhang, H., Yushkevich, P.A., Alexander, D.C., Gee, J.C.: Deformable registration of diffusion tensor MR images with explicit orientation optimization. Medical Image Analysis 10(5), 764–785 (2006)
14. Artaechevarria, X., Munoz-Barrutia, A., Ortiz-de Solorzano, C.: Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Transactions on Medical Imaging 28(8), 1266–1277 (2009)
15. Lorenzi, M., Pennec, X.: Efficient parallel transport of deformations in time series of images: From schilds to pole ladder. Journal of Mathematical Imaging and Vision, 1–13 (October 2013)

The 4D Hyperspherical Diffusion Wavelet: A New Method for the Detection of Localized Anatomical Variation

Ameer Pasha Hosseinbor[1], Won Hwa Kim[1], Nagesh Adluru[1], Amit Acharya[2], Houri K. Vorperian[1], and Moo K. Chung[1]

[1] University of Wisconsin-Madison, USA
[2] Marshfield Clinic, USA
hosseinbor@wisc.edu

Abstract. Recently, the HyperSPHARM algorithm was proposed to parameterize multiple disjoint objects in a holistic manner using the 4D hyperspherical harmonics. The HyperSPHARM coefficients are global; they cannot be used to directly infer localized variations in signal. In this paper, we present a unified wavelet framework that links HyperSPHARM to the diffusion wavelet transform. Specifically, we will show that the HyperSPHARM basis forms a subset of a wavelet-based multiscale representation of surface-based signals. This wavelet, termed the hyperspherical diffusion wavelet, is a consequence of the equivalence of isotropic heat diffusion smoothing and the diffusion wavelet transform on the hypersphere. Our framework allows for the statistical inference of highly localized anatomical changes, which we demonstrate in the first-ever developmental study on the hyoid bone investigating gender and age effects. We also show that the hyperspherical wavelet successfully picks up group-wise differences that are barely detectable using SPHARM.

1 Introduction

Studying and quantifying the development of anatomical structures over time is important in medical image analysis. Anatomical development tends to exhibit highly localized, complex growth [9]. Unfortunately, existing surface-based morphometric techniques are based on global bases, and thus are unable to detect subtle localized anatomical variations. For anatomical developmental studies, there is then a real need for surface-based approaches with localization power.

Recently, the HyperSPHARM algorithm [7] was proposed to parameterize multiple disjoint structures (e.g. hyoid bone) in a holistic manner. The underlying idea behind HyperSPHARM is to stereographically project n-dimensional data onto the $(n+1)$-dimensional hypersphere and subsequenly parameterize it with the $(n+1)$-dimensional hyperspherical harmonics (HSH). As with SPHARM [5,8], the HyperSPHARM coefficients are global, so if they exhibit statistical differences, interpreting which anatomical regions contribute to these variations is difficult. Consequently, the HyperSPHARM coefficients cannot be used directly

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 65–72, 2014.

to infer localized variations in signal. Although this fact may seem to render HyperSPHARM as a purely global inference algorithm, HyperSPHARM is actually a feature of wavelet localization.

In this paper, a unified wavelet framework is developed that links HyperSPHARM to the diffusion wavelet transform [4]. Specifically, we will show that the HyperSPHARM basis forms a subset of a wavelet-based multi-scale representation of surface-based signals. We will derive this wavelet, which we term the hyperspherical diffusion wavelet. Our framework allows for the statistical inference of highly localized anatomical changes, which we demonstrate in a developmental study on the hyoid bone investigating gender and age effects. We will also show that the hyperspherical wavelet outperforms SPHARM in detecting group-wise differences.

2 Theory

In this section, we will briefly review HyperSPHARM, before deriving the hyperspherical wavelet and its corresponding coefficients.

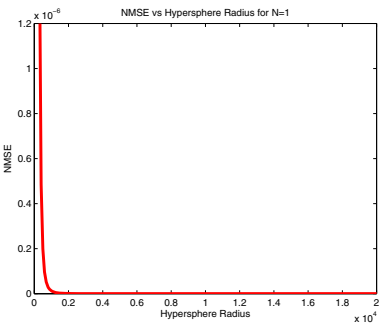


Fig. 1. NMSE versus hypersphere radius for N=1 HSH recon of hyoid template

4D Hyperspherical Harmonics
Consider the 4D unit hypersphere S^3 existing in $\mathbb{R}^4$. The Laplace-Beltrami operator on S^3 is defined as $\Delta_{S^3} = \frac{1}{\sin^2\beta}\frac{\partial}{\partial\beta}\sin^2\beta\frac{\partial}{\partial\beta} + \frac{1}{\sin^2\beta}\Delta_{S^2}$, where Δ_{S^2} is the Laplace-Beltrami operator on the unit sphere S^2. The eigenfuctions of Δ_{S^3} are the 4D HSH $Z_{nl}^m(\beta,\theta,\phi)$: $\Delta_{S^3}Z_{nl}^m = -l(l+2)Z_{nl}^m$, and we refer the reader to [7] for their exact functional form. The hyperspherical angles (β,θ,ϕ) obey $(\beta\in[0,\pi],\theta\in[0,\pi],\phi\in[0,2\pi])$, and the three integers (n,l,m) obey the conditions $n=0,1,2,...,\ 0\le l\le n$, and $-l\le m\le l$. The 4D HSH form an orthonormal basis on S^3.

Isotropic Heat Diffusion Smoothing on 4D Hypersphere
Consider an arbitrary 3D manifold $\mathcal{M}\subset\mathbb{R}^3$ defined by surface coordinates $\mathbf{q}=(q,\theta,\phi)$, and some real-valued functional measurement $f(\mathbf{q})$ defined on the manifold. The manifold $\mathcal{M}$ can be either multiple disjoint components such as the hyoid bone, or a single connected component. We stereographically project the 3D manifold onto a 4D hypersphere of radius r_o in $\mathbb{R}^4$, whose coordinates are denoted by the vector $\mathbf{u}=(\beta,\theta,\phi)$. Consequently, the functional measurement f exists along the surface of the 4D hypersphere. Note that the measurement $f(\mathbf{q})$ on $\mathcal{M}$ is equivalent to its corresponding projection $f(\mathbf{u})$ on the hypersphere.

We assume that $f(\mathbf{u})$ is square-integrable along the surface of the hypersphere. According to Fourier analysis, any square-integrable function defined on

a sphere can be expanded in terms of the spherical harmonics. Thus, $f(\mathbf{u})$ can be expanded in terms of the 4D HSH:

$$f(\mathbf{u}) = \sum_{n=0}^{N} \sum_{l=0}^{n} \sum_{m=-l}^{l} C_{nlm} Z_{nl}^{m}(\mathbf{u}), \tag{1}$$

where N is the truncation order of the HSH expansion. Eq. (1) is simply the HyperSPHARM basis.

Now lets have $f(\mathbf{u})$ undergo isotropic heat diffusion smoothing. We want to determine the function $K(\mathbf{u}, t)$ that describes the variation of $f(\mathbf{u})$ with respect to smoothing parameter t. Naturally, when no smoothing is applied, i.e. $t = 0$, we have $K(\mathbf{u}, 0) = f(\mathbf{u})$. The function $K(\mathbf{u}, t)$ is then a solution to the isotropic heat equation on the 4D hypersphere, subject to the aforementioned initial condition:

$$\frac{\partial}{\partial t} K(\mathbf{u}, t) - \Delta_{S^3} K(\mathbf{u}, t) = 0, \quad K(\mathbf{u}, t = 0) = f(\mathbf{u}) \tag{2}$$

Eq. (2) can be solved analytically by employing an ansatz solution of the form $K(\mathbf{u}, t) = \sum_{n=0}^{\infty} \sum_{l=0}^{n} \sum_{m=-l}^{l} C_{nlm} h_{nlm}(t) Z_{nl}^{m}(\mathbf{u})$, where $h_{nlm}(t)$ is the smoothing term. Upon substituting the ansatz solution into (2), we determine the smoothing term to be $h_l(t) = b_l e^{-l(l+2)t}$. Hence, the solution to Eq. (2) is

$$K(\mathbf{u}, t) = \sum_{n=0}^{N} \sum_{l=0}^{n} \sum_{m=-l}^{l} C_{nlm} e^{-l(l+2)t} Z_{nl}^{m}(\mathbf{u}), \tag{3}$$

where all constants are absorbed into C_{nlm}.

Connection to Diffusion Wavelet

Diffusion wavelets are a multi-scale framework for the analysis of functions on manifolds and graphs [1,6]. Consider the eigenfunctions ψ_j and eigenvalues λ_j on an arbitrary d-dimensional manifold $\mathcal{M}_d$, which satisfy $\Im \psi_j = \lambda_j \psi_j$ for some self-adjoint operator $\Im$ defined on $\mathcal{M}_d$. Following the notations in [1,6], the diffusion wavelet $W_{t,p}(s)$ at position p and scale t characterizing the manifold $\mathcal{M}_d$ is given by

$$W_{t,p}(s) = \sum_{j=0}^{k} g(\lambda_j, t) \psi_j(p) \psi_j(s), \tag{4}$$

where g is some scaling function. The diffusion wavelet coefficients of a given function $\epsilon(s)$ existing on the manifold $\mathcal{M}_d$ is given by the inner product of the wavelets and the given function:

$$\langle W_{t,p}, \epsilon \rangle = \int_M W_{t,p}(s) \epsilon(s) ds \tag{5}$$

If the manifold $\mathcal{M}_d$ is taken to be the 4D hypersphere, then $\Im$ is the Laplace-Beltrami operator on S^3; ψ_j is the 4D HSH basis Z_{nl}^m; and $\lambda_j = -l(l+2)$. Then Eq. (4) becomes

$$\boxed{W_{t,\mathbf{u}}(\mathbf{v}) = \sum_{n=0}^{N} \sum_{l=0}^{n} \sum_{m=-l}^{l} e^{-l(l+2)t} Z_{nl}^m(\mathbf{u}) Z_{nl}^m(\mathbf{v}),} \tag{6}$$

where we have taken $g(\lambda_j, t) = e^{-l(l+2)t}$. Eq. (6) is the 4D hyperspherical diffusion wavelet. Substituting Eqs. (1) and (6) into Eq. (5) gives the hyperspherical wavelet coefficients of any functional measurement f existing on the 4D hypersphere:

$$\boxed{\langle W_{t,\mathbf{u}}, f \rangle = \sum_{n=0}^{N} \sum_{l=0}^{n} \sum_{m=-l}^{l} C_{nlm} e^{-l(l+2)t} Z_{nl}^m(\mathbf{u})} \tag{7}$$

Eq. (7) is equivalent to Eq. (3), i.e. the hyperspherical wavelet coefficients are simply the functional measurement at different smoothing scales, which indicates that isotropic heat diffusion smoothing and the diffusion wavelet transform are identical operations on the hypersphere. It should be noted that a similar 3D analysis will result in the SPHARM wavelet.

3 Application

Numerical Implementation
The numerical implementation follows that of the HyperSPHARM algorithm. The task at hand is to estimate the HSH coefficients C_{nlm} in Eq. (1) for the functional measurement $f = (x^1, x^2, x^3)$, where (x^1, x^2, x^3) are the surface coordinates of the 3D manifold $\mathcal{M}$.

Suppose the manifold $\mathcal{M}$ comprises M mesh vertices, and let $\Omega_j = (\beta_j, \theta_j, \phi_j)$ denote the hyperspherical angles at the j-th mesh vertex. Denote $\mathbf{x}^i$ as the M x 1 vector representing each x^i's M vertices, $\mathbf{C}^i$ the Q x 1 vector of unknown expansion coefficients C_{nlm}^i for each x^i, and $\mathbf{A}$ the M x Q matrix constructed with the HSH basis

$$\mathbf{A} = \begin{pmatrix} Z_{00}^0(\Omega_1) & Z_{10}^0(\Omega_1) & Z_{11}^{-1}(\Omega_1) & Z_{11}^0(\Omega_1) & \cdots & Z_{NN}^N(\Omega_1) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_{00}^0(\Omega_M) & Z_{10}^0(\Omega_M) & Z_{11}^{-1}(\Omega_M) & Z_{11}^0(\Omega_M) & \cdots & Z_{NN}^N(\Omega_M) \end{pmatrix}.$$

Thus, the general linear system representing Eq. (1) is described by $\mathbf{x}^i = \mathbf{A}\mathbf{C}^i$, and is solved via linear least squares, yielding $\widehat{\mathbf{C}^i} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{x}^i$. Once the HyperSPHARM coefficients have been estimated, heat diffusion smoothing is applied to obtain the hyperspherical wavelet coefficients given by Eq. (7).

CT Imaging Data and Preprocessng

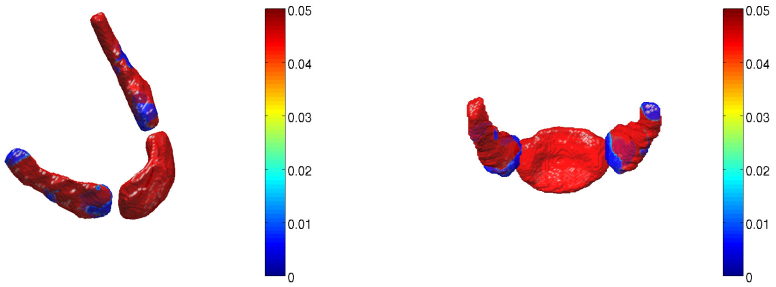
The study consists of CT images of 70 normal subjects (33 female and 37 male), whose age range is between 0 and 20 years. Subjects are binned into three age categories: ages between 0 and 6 years (group I), 7 and 12 years (group II), and 13 and 19 years (group III). There are 26, 14, and 30 subjects in groups I, II, and III, respectively. Using this dataset, we seek to address two issues: 1) whether there are any localized hyoid bone growth spurts between these age groups and 2) whether there are any gender differences in the hyoid bone.

The hyoid bone was segmented manually. Correspondence for SPHARM and HyperSPHARM was established in a similar manner as proposed in [3]. The 70 subjects were first affinely aligned so to remove the overall size variablity. Since some subjects may have a larger hyoid bone than others, it is necessary to remove the global size differences in local shape modeling. For this reason, diffeomorphic nonlinear image registration was then performed on the affinely registered template using Advanced Normalization Tools (ANTS) [2]. SPHARM and HyperSPHARM are then used to further register the surfaces via surface flattening and stereographic projection, respectively. Please note this approach avoids the surface alignment done by coinciding the first order ellipsoid meridian and equator proposed in [5].

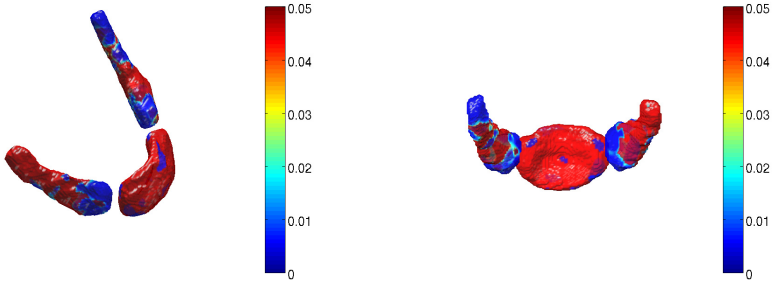
The HSH truncation order was $N = 1$ and hypersphere radius $r_o = 12000$. The appropriate radius was determined by plotting the mean squared error as a function of radius and selecting the radius that minimized it (Fig. 1). The SPHARM truncation order was $L = 20$. The appropriate wavelet scales t were determined using cluster size inference. The hyperspherical and SPHARM wavelet coefficients are estimated for each vertex at scales $t = [0.01\ 0.05]$ for gender and $t = 0.005$ for age. The SPHARM estimation is special generate case of the SPHARM wavelet at $t = 0$. Hotelling's T^2 test was then carried out at the voxel level at .05 significance level for group analysis with respect to age and gender. The resulting p-value map was corrected for multiple comparisons across all vertices using the false discovery rate (FDR) method.

Hotelling T^2 Statistical Results and Discussion

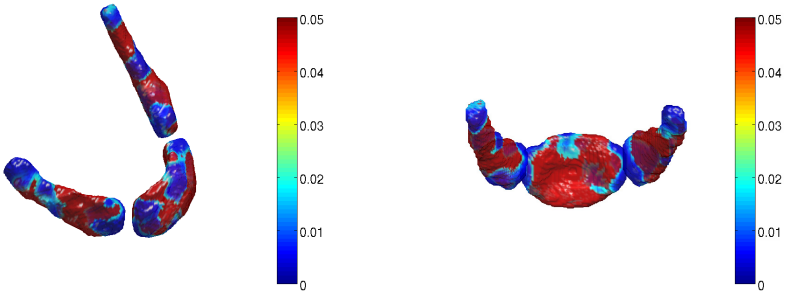
Only results related to gender and age groups I vs. II are presented. Figs. 2 and 3 summarize the results of our analysis using hyperspherical/SPHARM wavelets and SPHARM, with non-red regions indicating statistical significance. All three methods detect significant gender differences and growth spurts at several regions along the right and left hyoid bones and near the regions that connect the disconnected hyoid bones. SPHARM, however, detects no significant gender and age effects in the middle hyoid bone, unlike the hyperspherical wavelet. The SPHARM wavelet does detect significant gender differences in a few areas along the middle hyoid bone, but no age effects. For both age and gender, the hyperspherical wavelet had the largest number of significant vertices, followed by SPHARM wavelet, and then SPHARM. For gender, the hyperspherical wavelet has a total of 8575 statistically significant vertices, whereas SPHARM wavelet has 6384 and SPHARM 2928. For age, the hyperspherical wavelet detects 5394 statistically significant vertices, followed by SPHARM wavelet with 5330, and SPHARM with 4854.



SPHARM: Gender

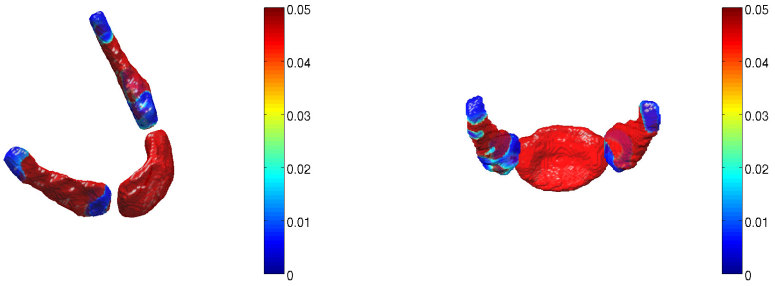


SPHARM Wavelet: Gender

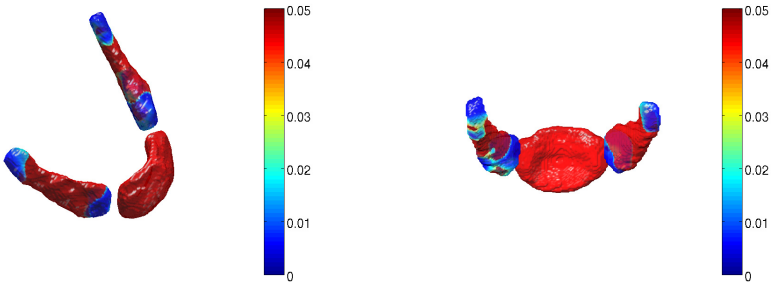


Hyperspherical Wavelet: Gender

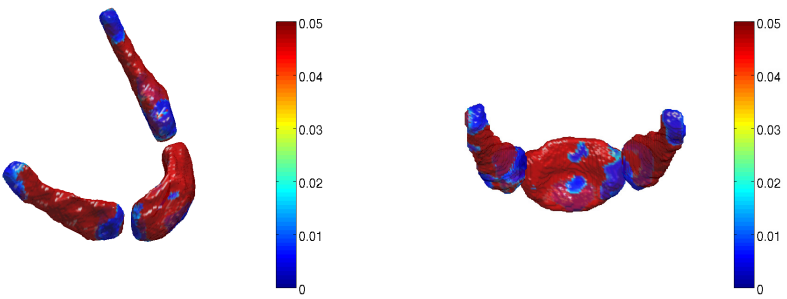
Fig. 2. Testing for gender differences. p-values after FDR correction (i.e. q-value), projected back onto hyoid bone template. For hyperspherical wavelet, q-value < 0.028 corresponds to significance; for SPHARM wavelet, q-value < 0.029 (LH), < 0.005 (MH), and < 0.019 (RH); for SPHARM, q-value < 0.011 (LH) and < 0.014 (RH).



SPHARM: Age Groups I & II



SPHARM Wavelet: Age Groups I & II



Hyperspherical Wavelet: Age Groups I & II

Fig. 3. Testing for age effects. p-values after FDR correction (i.e. q-value), projected back onto hyoid bone template. For hyperspherical wavelet, q-value < 0.014 corresponds to significance; for SPHARM wavelet, q-value < 0.028 (LH) and < 0.016 (RH); for SPHARM, q-value < 0.026 (LH) and < 0.014 (RH).

The hyperspherical diffusion wavelet's outperformance of SPHARM is due to the wavelet being a local basis whereas SPHARM employs a global basis. The wavelet's inherent localization power, therefore, enables it to infer localized shape variations much better than globally-based methods like SPHARM. SPHARM wavelet's outperformance of SPHARM is for the same reason.

4 Conclusion

In this paper, we have introduced the hyperspherical diffusion wavelet, which allows for the statistical detection of highly localized variations in anatomical morphology. It was used in the first ever developmental study on the hyoid bone, and subsequent statistical testing on the wavelet coefficients revealed localized gender differences and growth spurts in the hyoid bone. We also showed that our framework is more sensitive in signal detection, outperforming both SPHARM wavelet and SPHARM in the discernment of group-wise differences.

Acknowledgements. This research was supported by NIH grants R01DC006282, P30HD003352, and 5T15LM007359, and Vilas Associate Award from UW-Madison.

References

1. Antoine, J.P., Rosca, D., Vandergheynst, P.: Wavelet transform on manifolds: old and new approaches. Applied and Computational Harmonic Analysis 28, 189–202 (2010)
2. Avants, B., Epstein, C., Grossman, M., Gee, J.: Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis 12, 26–41 (2008)
3. Chung, M.K., Dalton, K.M., Shen, L., Evans, A.C., Davidson, R.J.: Weighted Fourier series representation and its application to quantifying the amount of gray matter. IEEE Transac. Med. Imaging 26, 566–581 (2007)
4. Coifman, R., Maggioni, M.: Diffusion wavelets. Applied and Computational Harmonic Analysis 21, 53–94 (2006)
5. Gerig, G., Styner, M., Jones, D., Weinberger, D., Lieberman, J.: Shape analysis of brain ventricles using spharm. In: MMBIA, pp. 171–178 (2001)
6. Hammond, D.K., Vandergheynst, P., Gribonval, R.: Wavelets on graphs via spectral graph theory. Applied and Computational Harmonic Analysis 30, 129–150 (2011)
7. Hosseinbor, A.P., Chung, M.K., Schaefer, S.M., van Reekum, C.M., Peschke-Schmitz, L., Sutterer, M., Alexander, A.L., Davidson, R.J.: 4D hyperspherical harmonic (HyperSPHARM) representation of multiple disconnected brain subcortical structures. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part I. LNCS, vol. 8149, pp. 598–605. Springer, Heidelberg (2013)
8. Shen, L., Ford, J., Makedon, F., Saykin, A.: surface-based approach for classification of 3d neuroanatomical structures. Intelligent Data Analysis 8, 519–542 (2004)
9. Vorperian, H.K., Wang, S., Schimek, E.M., Durtschi, R.B., Kent, R.D., Gentry, L.R., Chung, M.K.: Developmental sexual dimorphism of the oral and pharyngeal portions of the vocal tract: an imaging study. JSLHR 54, 995–1010 (2011)

Co-occurrence of Local Anisotropic Gradient Orientations (CoLlAGe): Distinguishing Tumor Confounders and Molecular Subtypes on MRI

Prateek Prasanna*, Pallavi Tiwari*, and Anant Madabhushi

Department of Biomedical Engineering, Case Western Reserve University, USA
pallavi.tiwari@case.edu

Abstract. We introduce a novel *biologically inspired* feature descriptor, Co-occurrence of Local Anisotropic Gradient Orientations (CoLlAGe), that captures higher order co-occurrence patterns of local gradient tensors at a pixel level to distinguish disease phenotypes that have similar morphologic appearances. A number of pathologies (e.g. subtypes of breast cancer) have different histologic phenotypes but similar radiographic appearances. While texture features have been previously employed for distinguishing subtly different pathologies, they attempt to capture differences in global intensity patterns. In this paper we attempt to model CoLlAGe to identify higher order co-occurrence patterns of gradient tensors at a pixel level. The assumption behind this new feature is that different pathologies, even though they may have very similar overall texture and appearance on imaging, at a local scale, will have different co-occurring patterns with respect to gradient orientations. We demonstrate the utility of CoLlAGe in distinguishing two subtly different types of pathologies on MRI in the context of brain tumors and breast cancer. In the first problem, we look at CoLlAGe for distinguishing radiation effects from recurrent brain tumors over a cohort of 40 studies, and in the second, discriminating different molecular subtypes of breast cancer over a cohort of 73 studies. For both these challenging cohorts, CoLlAGe was found to have significantly improved classification performance, as compared to the traditional texture features such as Haralick, Gabor, local binary patterns, and histogram of gradients.

1 Introduction

A number of pathologies have different histologic phenotypes but similar radiographic appearances. One such instance is the problem of discriminating fibroadenoma (FA), a benign breast tumor from triple negative (TN), an aggressive breast cancer [1]. Both FA and TN have distinct cellular and architectural

* P. Prasanna and P. Tiwari are joint first authors. Research was supported by R01CA136535-01, R01CA140772-01, R21CA167811-01, R01DK098503-02, PC120857, the QED award, the Ohio Third Frontier Technology development Grant, and the Coulter Translational Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 73–80, 2014.

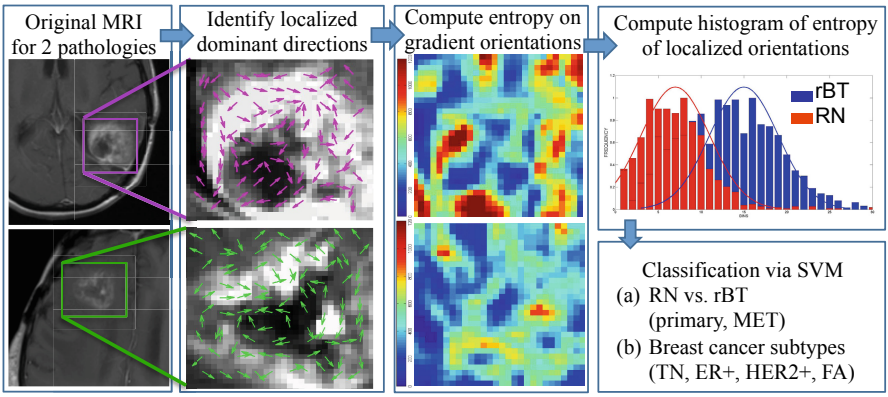


Fig. 1. Overview of CoLlAGe and overall workflow. The 1st module involves computing localized gradient orientations, while in the 2nd module entropy of localized gradient orientations is computed for every pixel. A histogram of entropy values is then aggregated for every pixel and is subsequently used for classification.

arrangements when examined on a pathology slide under a microscope. However they have very similar morphologic appearances on MRI. In this paper, we present a new *biologically inspired* feature descriptor, Co-occurrence of Local Anisotropic Gradient Orientations (CoLlAGe), that captures higher order co-occurrence patterns of local gradient tensors at a pixel level to distinguish disease phenotypes that have similar morphologic appearances. While texture features have emerged as a popular way of characterizing and distinguishing subtly differing pathologies, these operators typically tend to capture global textural patterns. One such class of texture features are grey-level co-occurrence matrix (GLCM) [2] and Gabor steerable filters. These texture descriptors involve computing global relationships between pixels by averaging responses to various filter operators within the neighborhood of a single global descriptor. While some texture features can provide pixel-level responses (e.g. local binary patterns (LBP)) [3], these filters are often employed to provide pixel level or patch based classification. LBP, unlike GLCM, provides a signature for every pixel by capturing localized intensity variations across the pixel. However, LBP, is highly dependent on the radius parameter, which is critical in extracting local patterns. Additionally, both global and per-pixel texture representations are based on intensity variations and are domain agnostic. When examined on a histopathology slide under a microscope at a high magnification, the differences between subtly different classes may be manifested in differently oriented nuclei, lymphocytes, and/or glands. The differences in histologic architecture, which are no doubt reflected on the imaging, hence need a new class of features to capture subtle differences in image patterns on a local scale.

The rationale behind our approach, CoLlAGe, is that even though overall the global textural patterns or even the filter responses at a majority of pixel

locations might be similar between two differing pathologies (e.g. FA versus TN), the organization of local gradients may differ across classes and will be relatively consistent within a class.

2 Previous Work and Novel Contributions

A popular texture descriptor that captures orientation variations is histogram of gradient orientations (HoG) [4]. HoG computes a global patch based signature by computing histogram distribution of orientations computed on a per pixel basis. However HoG, similar to the other texture descriptors, is domain agnostic and is not designed to capture localized per voxel texture characteristics depicted on imaging. A variant of HoG, called co-occurrence of histogram of gradient orientations (Co-HoG) was recently presented by Watanabe et al. [5] where a high-dimensional feature vector was computed for every pixel by accumulating values from co-occurrence matrix computed on gradient orientations for pedestrian detection on a per-pixel basis. However, the approach in [5], (a) did not capture localized variations across neighboring orientations, (b) is susceptible to "curse of dimensionality" (due to a high dimensional feature space), and (c) similar to its counterpart HoG, is domain agnostic.

Recently, deep learning (DL) has emerged as a powerful tool for learning alternative representations of data for improved classification [6]. DL approaches train multiple convolution layers on a large annotated dataset to learn abstract but useful patterns between classes. Although DL has shown tremendous promise in identifying complex differentiating patterns across diseases, the identified features are not intuitive and cannot be used to understand the underlying disease characteristics. Additionally, DL strategies require large annotated training dataset to obtain meaningful results.

CoLlAGe, on the other hand, is designed to be "biologically intuitive". Firstly, CoLlAGe captures neighborhood orientation variation via a localized gradient tensor field that may reflect the underlying cellular arrangement of the phenotype on imaging. Secondly, CoLlAGe computes co-occurrence matrix on localized gradient tensors to capture co-occurring patterns of orientation disorder in a localized fashion. While co-occurrence matrices are commonly used to describe image texture, to our knowledge, this is the first attempt at employing co-occurrences on localized gradient orientations to capture underlying orientation variations on imaging.

We demonstrate the utility of CoLlAGe in the context of two problems involving brain tumors and breast cancer. The first application involves evaluating radiation therapy response for distinguishing radiation necrosis (RN), a radiation induced effect, from recurrent brain tumors (rBT) for primary and metastatic (MET) brain tumors [7]. The second application involves identifying phenotypic imaging signatures of molecular sub-types of breast cancer: triple negative (TN), estrogen receptor positive (ER+), human epidermal growth factor receptor positive (HER2+), and benign fibroadenoma (FA) on dynamic contrast enhanced (DCE)-MRI [1].

3 Co-occurrence of Localized Gradient Orientations

A region of interest (ROI) on an MRI volume is defined as, $\mathcal{C} = (C, f)$, where $f(c)$ is the associated intensity at every pixel c on a 3D grid C. Computation of CoLlAGe for every $c \in C$ involves following steps,

1. **Calculation of gradient orientations for every pixel:** For every $c \in C$, gradients along the X and Y directions are computed as, $\nabla f(c) = \frac{\partial f(c)}{\partial X}\hat{i} + \frac{\partial f(c)}{\partial Y}\hat{j}$. Here, $\frac{\partial f(c)}{\partial X}$ and $\frac{\partial f(c)}{\partial Y}$ are the gradient magnitudes along the X and the Y axes respectively denoted by $\partial f_X(c)$ and $\partial f_Y(c)$. The gradient orientation θ of every $c \in C$ is then calculated as $\theta(c) = \tan^{-1}\frac{\partial f_Y(c)}{\partial f_X(c)}$.
2. **Computing local dominant orientations via principal component analysis (PCA):** A $\mathcal{N}\times\mathcal{N}$ window centered around every $c \in C$ is selected. We then compute $\partial f_X(c_k)$ and $\partial f_Y(c_k)$, $k \in \{1, 2, \ldots, \mathcal{N}^2\}$. The vector gradient matrix $\overrightarrow{\mathbf{F}}$ associated with every c is given by $\overrightarrow{\mathbf{F}} = [\overrightarrow{\partial f}_X(c_k)\ \ \overrightarrow{\partial f}_Y(c_k)]$, where $[\overrightarrow{\partial f}_X(c_k)\ \ \overrightarrow{\partial f}_Y(c_k)]$, $k \in \{1, 2, \ldots, \mathcal{N}^2\}$ is the matrix of gradient vectors in the X and Y directions for every c_k. The most significant orientation for each pixel c_k within $\mathcal{N} \times \mathcal{N}$ is obtained by performing principal component analysis on $\overrightarrow{\mathbf{F}}$. The dominant principal components in X and Y directions are obtained as r_X^k and r_Y^k for every $k \in \{1, 2, \ldots, \mathcal{N}^2\}$. The most significant orientation for every c_k is then calculated as $\phi(c) = \tan^{-1}\frac{r_Y^k}{r_X^k}$.
3. **Calculation of second-order statistics for most significant orientations:** The objects of interest for calculating CoLlAGe features are the co-occurring directions given by discretization of the dominant orientation $\bar{\phi}(c_k)$ for every pixel c, such that $\phi(c_k) = \omega \times \text{ceil}(\frac{\bar{\phi}(c_k)}{\omega})$, where ω is a discretization factor.
   An $N \times N$ co-occurrence matrix $\mathcal{M}$ subsequently captures orientation pairs between pixels which co-occur in the neighborhood $\mathcal{W}_i$, such that,

$$\mathcal{M}_{\mathcal{W}_i}(p, q) = \sum_{c_j, c_k}^{\mathcal{W}_i} \sum_{p,q=1}^{N} \begin{cases} 1, & \text{if } \phi(c_j)=p \text{ and } \phi(c_k)=q \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

   where $N = \frac{360}{\omega}$ is the number of discrete angular bins. Entropy measure, $\mathcal{E}(c)$ is then computed from every co-occurrence matrix on every c as,

$$\mathcal{E}(c) = \sum_{p,q} -\mathcal{M}(p, q) log(\mathcal{M}(p, q)). \tag{2}$$

4. A **histogram of** $\mathcal{E}$ is computed by aggregating $\mathcal{E}(c_k)$, $k \in \{1, \ldots, |C|\}$, where $|.|$ is the cardinality of set C. The entropy histogram is divided into bin size v, optimized on the training set via grid search optimization.
   A CoLlAGe feature vector, $\mathcal{F}$ is then obtained for every $\mathcal{C}$ which consists of the binned histogram values in the form of v×1 vectors. $\mathcal{F}$ is then employed within a classifier for classification purposes.

Data: ROI volume $\mathcal{C}$
Result: *CoLlAGe* features, $\mathcal{F}$

begin
  for each pixel $c \in \mathcal{C}$, **do**
    Obtain gradients $\partial f_X(c)$ and $\partial f_Y(c)$ along X- and Y- axes ;
    Obtain gradient orientation $\theta(c) = \tan^{-1} \frac{\partial f_Y(c)}{\partial f_X(c)}$;
  end
  for each pixel $c \in \mathcal{C}$, **do**
    Compute gradient vectors $\overrightarrow{\partial f_X}(c_k)$ and $\overrightarrow{\partial f_Y}(c_k)$ in $\mathcal{N} \times \mathcal{N}$ neighborhood;
    Obtain localized gradient vector matrix $\overrightarrow{\mathcal{F}} = [\overrightarrow{\partial f_X}(c_k) \ \overrightarrow{\partial f_Y}(c_k)]$;
    Compute dominant orientation, $\phi(c_k)$, $k \in \{1, \ldots, \mathcal{N}^2\}$ via PCA;
  end
  Compute Co-occurrence matrix $\mathcal{M}$ from $\phi(c)$;
  Compute $\mathcal{E}(c)$ from $\mathcal{M}$;
  Obtain v×1 dimensional feature vector $\mathcal{F}$ from distribution of $\mathcal{E}$;
end

Algorithm 1. Computation of *CoLlAGe* features

4 Experimental Results and Description

4.1 Data Description

Dataset 1 comprised two cohorts of 20 primary (10 RN, and 10 rBT) and 20 MET (12 RN and 8 rBT) patient studies respectively. All the studies were retrospectively acquired with 3 Tesla Gadolinium-constrast (Gd-C) T1-w MRI), and were histologically confirmed on biopsy samples by an expert pathologist. Dataset 2 comprised DCE-MRI studies from 65 women with 73 breast lesions for whom pathology results and ER, PR, and HER2 results were available. Reference standard diagnosis was made by histopathologic examination of tissue obtained by either core biopsy sampling or lumpectomy. Of the 73 lesions, 9 were benign FA, 21 were TN, 18 were HER2+, and 25 were ER+.

4.2 Implementation Details

Figure 1 shows the work-flow of CoLlAGe, and its implementation in the context of clinical problems in brain tumors and breast cancer. We compared CoLlAGe against GLCM, Gabor, HoG, and LBP features, and evaluated their performance using a support vector machine (SVM) classifier [8] with a radial basis function (RBF) kernel. A 3-fold cross-validation strategy was employed and the performance of each of the texture descriptors was compared over different window, $\mathcal{N} \in \{3, 5, 7, 9\}$, and bin sizes, $v \in \{10, 20, 30, 40, 50\}$. $v = 30$, and $\mathcal{N} = 7$ were found to be optimal parameters across different descriptors and was employed for further evaluation. The average accuracy values were reported over 100 runs of 3-fold cross validation for both the cohorts. Wilcoxon's rank sum test [9] was performed to report statistical significance and corrected for multiple comparisons for the experiments performed for the two use-cases.

Table 1. Summary of features and feature parameters used in this work

Descriptor	#	Feature setting	Description
Haralick	26	$\mathcal{N} \in \{3, 5, 7, 9\}$	gray-level co-occurrence such as angular second moment, contrast and entropy
LBP	59	radius $R = 8$	Histogram of intensity variations within R
HoG	20	bin-size=18°	Histogram of gradient orientations
CoLlAGe	30	$\omega = 20$; $\mathcal{N} \in \{3, 5, 7, 9\}$; $v \in \{10, 20, 30, 40\}$	Entropy of localized gradients for $\mathcal{N} \times \mathcal{N}$

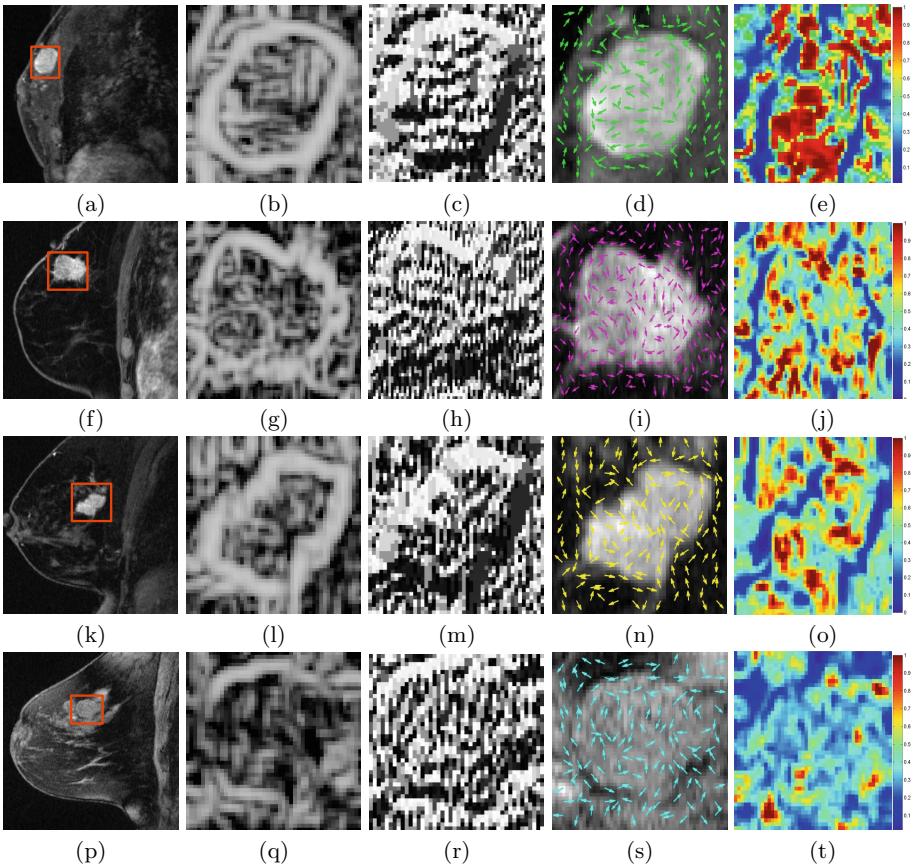


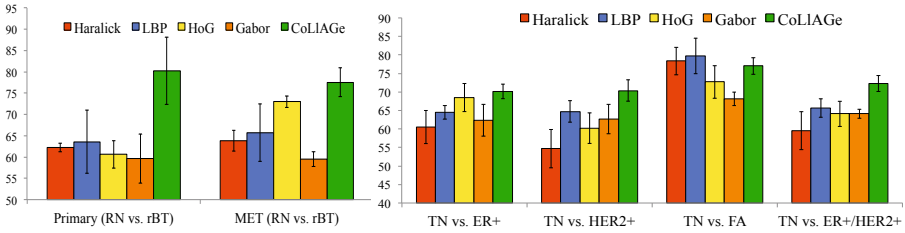
Fig. 2. Original DCE-MRI for ER+ (a), TN (f), HER2+ (k), and FA (p) with the lesion outlined in red. 2(b), (g), (l), (q) correspond to Haralick, while 2(c), (h), (m), (r) correspond to LBP representations of the lesion on ER+, TN, HER2+, and FA images. 2(d), (i), (n), (s) represent localized gradient orientations, while 2(e), (j), (o), (t) represent entropy heatmaps for the corresponding lesion on (a), (f), (k) and (p), where red represents higher while blue represents low entropy values.

4.3 Distinguishing RN from rBT for Primary and MET Patients

Figure 1 shows a representative RN, and rBT patient for primary brain tumor cohort. The orientations shown in magenta correspond to rBT, while the ones in green correspond to RN. The heatmaps represent entropy values obtained from the localized orientations on a per-pixel basis, where higher values of entropy are shown in red while lower values are shown in blue. It is interesting to note that the entropy values for rBT on a per pixel basis are substantially higher than those of RN suggesting orientation disorder in recurrent tumor. The histogram plots shown in red (RN), and blue (rBT) seem to suggest a clear separation between entropy distributions across the two morphologies. Figure 3 (a) demonstrates the quantitative results obtained for both primary as well as MET patients in distinguishing RN from rBT. For both the cohorts, CoLlAGe (80.25 $\pm$ 7.89% for primary cases, 77.55% $\pm$ 3.35 for MET) was found to significantly outperform Haralick (62.19% $\pm$ 0.99% for primary, 63.83% $\pm$ 2.42% for MET), Gabor (59.68% $\pm$ 5.8% for primary, 59.45% $\pm$ 1.73% for MET), LBP (63.63 $\pm$ 3.21% for primary, 65.75% $\pm$ 6.76% for MET), and HoG features (60.62 $\pm$ 3.21% for primary, 72.99% $\pm$ 1.35 for MET).

4.4 Distinguishing TN from Other Breast Cancer Subtypes

Figure 2 illustrates the qualitative comparison of CoLlAGe with the other texture descriptors, Haralick, LBP, and HoG in differentiating molecular subtypes, ER+ (2(a)), TN (2(f)), HER2+ (2(k)), and FA (2(p)) on DCE-MRI. Note the apparent differences in entropy heatmaps across 2(e), (j), (o), and (t) corresponding to ER+, TN, HER2+, and FA respectively. The most prominent difference is reported between FA, a benign condition, from ER+, a subtype of breast cancer. The results suggest that the orientations of cancer subtypes are more disordered than the benign condition. Similarly, ER+ reported overall higher entropy values than TN, and HER2+ cancer subtypes. The accuracy values averaged over 100 runs of 3 fold cross-validation for different feature descriptors are shown in Figure 3(b). It is interesting to note that although CoLlAGe significantly outperforms the other texture descriptors (Haralick, Gabor, LBP, and HoG) for



(a) (b)

Fig. 3. Mean accuracy values obtained for different feature descriptors (Haralick, LBP, HoG, Gabor, CoLlAGe) for (a) brain tumor, and (b) breast tumor datasets respectively.

distinguishing the more difficult disease subtypes (TN vs. HER2+, TN vs. ER+ and TN vs. ER+/HER2+), the results of CoLlAGe were comparable to Haralick descriptors for distinguishing TN vs. FA.

5 Concluding Remarks

In this work, we presented a new feature descriptor, Co-occurrence of Local Anisotropic Gradient Orientations (CoLlAGe), that captures higher order co-occurrence patterns of local gradient tensors at a pixel level to distinguish disease phenotypes that have similar morphologic appearances. We demonstrated the utility of CoLlAGe in identifying MRI phenotypes for clinically challenging problems in the context of breast and brain tumors for (a) distinguishing radiation necrosis, a treatment related effect from recurrent brain tumors over a cohort of 40 MRI studies, and (b) distinguishing breast cancer subtypes (ER+, HER2+, TN, and FA) over a cohort of 73 DCE-MRI studies. CoLlAGe was found to significantly outperform traditional texture descriptors such as Haralick, Gabor, local binary patterns, and histogram of gradients for the two use-cases, except for distinguishing FA from TN, where the results were found to be comparable to Haralick texture descriptors. In future work, we will seek to understand the correlation of CoLlAGE with pathologic correlates and evaluate its applicability across other disease sites such as prostate cancer.

References

1. Agner, S., et al.: Computerized image analysis for identifying triple-negative breast cancers and differentiating them from other molecular subtypes of breast cancer on dce-mri. Radiology (pre-print, 2014)
2. Haralick, R.M., et al.: Textural Features for Image Classification. Systems, Man and Cybernetics (6), 610–621 (1973)
3. Ojala, T., et al.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. PAMI 24(7), 971–987 (2002)
4. Dalal, N., Triggs, B.: Histograms of Oriented Gradients forHuman Detection. In: CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
5. Pang, Y., et al.: Robust Cohog Feature Extraction in Human-Centered Image/Video Management System. Systems, Man, and Cybernetics 42(2), 458–468 (2012)
6. Bengio, Y., Courville, A.C., Vincent, P.: Unsupervised feature learning and deep learning: A review and new perspectives. CoRR (2012)
7. Tiwari, P., et al.: Texture descriptors to distinguish radiation necrosis from recurrent brain tumors on multi-parametric mri. In: SPIE, pp. 90352B–90352B (2014)
8. Furey, T.S., et al.: Svm Classification and Validation of Cancer Tissue Samples using Microarray Expression Data. Bioinformatics 16(10), 906–914 (2000)
9. Wilcoxon, F., Wilcox, R.A.: Some Rapid Approximate Statistical Procedures. Lederle Laboratories (1964)

Automatic Clustering and Thickness Measurement of Anatomical Variants of the Human Perirhinal Cortex

Long Xie[1], John Pluta[1], Hongzhi Wang[5], Sandhitsu R. Das[1], Lauren Mancuso[2,3], Dasha Kliot[2,3], Brian B. Avants[1], Song-Lin Ding[4], David A. Wolk[2,3], and Paul A. Yushkevich[1]

[1] Penn Image Computing and Science Laboratory, Department of Radiology, University of Pennsylvania, Philadelphia, USA
[2] Penn Memory Center, University of Pennsylvania, Philadelphia, USA
[3] Department of Neurology, University of Pennsylvania, Philadelphia, USA
[4] Allen Institute for Brain Science, Seattle, USA
[5] IBM Almaden Research Center, San Jose, USA

Abstract. The entorhinal cortex (ERC) and the perirhinal cortex (PRC) are subregions of the medial temporal lobe (MTL) that play important roles in episodic memory representations, as well as serving as a conduit between other neocortical areas and the hippocampus. They are also the sites where neuronal damage first occurs in Alzheimer's disease (AD). The ability to automatically quantify the volume and thickness of the ERC and PRC is desirable because these localized measures can potentially serve as better imaging biomarkers for AD and other neurodegenerative diseases. However, large anatomical variation in the PRC makes it a challenging area for analysis. In order to address this problem, we propose an automatic segmentation, clustering, and thickness measurement approach that explicitly accounts for anatomical variation. The approach is targeted to highly anisotropic ($0.4 \times 0.4 \times 2.0 mm^3$) T2-weighted MRI scans that are preferred by many authors for detailed imaging of the MTL, but which pose challenges for segmentation and shape analysis. After automatically labeling MTL substructures using multi-atlas segmentation, our method clusters subjects into groups based on the shape of the PRC, constructs unbiased population templates for each group, and uses the smooth surface representations obtained during template construction to extract regional thickness measurements in the space of each subject. The proposed thickness measures are evaluated in the context of discrimination between patients with Mild Cognitive Impairment (MCI) and normal controls (NC).

1 Introduction

Quantification of the volume and thickness of ERC, PRC and other MTL cortical subregions from *in vivo* MRI has been increasingly pursued because these structures play important roles in episodic memory models [1] and are the earliest sites affected by AD pathology [2]. However, the PRC exhibits large anatomical variability, which complicates quantitative analysis [3]. By examining a large sample of autopsy brains, Ding *et al.* [4] conclude that three main variants of the PRC exist, defined by morphology of the collateral sulcus (CS): 1) continuous CS; 2) discontinuous CS with

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 81–88, 2014.

anterior CS shorter than the posterior; 3) discontinuous CS with anterior CS longer than the posterior. Failure to account for this variability can degrade the accuracy of morphometric analysis and reduce the utility of PRC as an imaging biomarker. This paper provides a novel approach for automatically quantifying the thickness of MTL substructures while explicitly accounting for anatomical variability.

Typically, the first step in quantitative MRI analysis is to segment the structures of interest, preferably automatically. However, little work on automatic segmentation of the PRC has been reported in the literature [5]. In this paper, we use the multi-atlas approach [6] in conjunction with a set of expert-labeled atlases that include labels for the ERC, PRC (further partitioned into Brodmann areas BA35 and BA36) as well as the hippocampal subfields (cornu ammonis, dentate gyrus and subiculum) to perform automatic segmentation. The method takes T1-weighted whole-brain scan ($1mm^3$ isotropic resolution) as well as a specialized anisotropic oblique coronal T2-weighted scan of the MTL ($0.4x0.4x2mm^3$ resolution) as inputs, and outputs a multi-label segmentation that has the same resolution as the T2-weighted image. The T2-weighted MRI has high in-plane resolution that allows substructures in the hippocampal region to be distinguished visually in the way that $1mm^3$ isotropic T1-weighted MRI cannot. Similar T2-weighted MRI scans have been used for manual segmentation of MTL substructures by several authors, e.g. [7,8].

Regional thickness measurements are often preferred to volume in morphometric studies of cortical structures like ERC and PRC because 1) they capture localized changes and 2) they are more robust to the variability of the locations of the boundaries in the automatic segmentation. While there is substantial prior work on measuring cortical thickness in MRI [9,10], most approaches do not provide a specific PRC thickness measurement. The notable exception is [5], who use a probabilistic template derived from *postmortem* MRI to label and measure the thickness of the PRC in the *in vivo* MRI. However, this single-template approach does not account for the anatomical variability described by Ding *et al.* [4]. In this paper, we propose a thickness measurement pipeline that attempts to automatically discover anatomical variants present in the population using a combination of deformable image registration and spectral clustering [11]. Our work is inspired by recent applications of clustering to atlas propagation and group-wise image registration [12], but is distinct in that clustering is applied to the output of multi-atlas segmentation rather than raw MRI data. The main contribution of this paper is introducing this concept in the analysis of PRC, which is the perfect application for this technique.

To demonstrate clinical utility, we evaluate our technique in a dataset from a research study of MCI, often conceptualized as a prodromal stage of AD, and normal aging. The proposed clustering-based approach yields stronger statistical power in discriminating the MCI patients from NC group than volumetric measurements as well as alternative thickness measures.

2 Materials

MRI scans of 83 participants (40 with diagnosis of MCI, 43 controls) from a research study conducted at the Penn Memory Center at the University of Pennsylvania were used to evaluate the proposed technique. Scans were acquired on a 3T Siemens Trio

scanner. MRI protocols include a T1-weighted (MPRAGE) $1mm^3$ isotropic whole-brain scan and a $0.4x0.4x2mm^3$ T2-weighted (TSE) scan with partial brain coverage and an oblique coronal slice orientation (Fig 1a,b).

Automatic segmentation for each subject was generated by applying the multi-atlas approach in [6] to the subject`s T1-weighted and T2-weighted scans (Fig 1c). The output segmentations, consist of 7 labels (cornu ammonis, dentate gyrus, subiculum, ERC, BA35, BA36 and CS), were then used for our proposed pipeline.

3 Method

Given the automatic segmentation, which has large step edge discontinuities in the MRI slice direction, our goal is to approximate it with a smooth surface mesh representation that is topologically consistent across all subjects sharing the same PRC anatomical subtype, and from which a regional thickness map can be extracted. Our proposed approach consists of three steps: 1) cluster subjects into groups based on their PRC anatomy; 2) build an unbiased population template for each group and generate a mesh in the template space; 3) warp the mesh back to the space of each subject and measure thickness for each vertex on the mesh. We treat each hemisphere independently throughout the analysis. The computational complexity of clustering and thickness analysis is negligible compared to multi-atlas segmentation. The detail of each step is described below.

3.1 Automatic Clustering of Anatomical Subtypes

Spectral clustering [11] is used to divide subjects into groups with similar PRC anatomy. Spectral clustering is a dimension reduction algorithm that projects the pairwise similarity relationship onto a lower-dimensional space in which anatomical variants can be easily separated using k-means clustering [13].

To compute the pairwise similarity matrix (denoted as S), we first perform pairwise registration between all the multi-label segmentations using ANTs affine and high-dimensional deformable algorithms [14]. The registration minimizes the sum of mean square intensity difference metrics computed separately for each label. Generalized Dice Similarity Coefficient (DSC) [15] is computed for labels BA35 and BA36 between the warped segmentation of subject i and the segmentation of subject j, and denoted as D_{ij}. The underlying assumption is that after registration, overlap between multi-label segmentations will be greater when the pair of subjects have the same anatomical variant of the PRC than when they have different variants. In order to have a symmetric measurement and also exaggerate the similarity value between subjects with similar PRC anatomy, we compute similarity between subject i and subject j as:

$$s_{ij} = \exp\{\frac{-[\log 0.5(D_{ij} + D_{ji})]^2}{2\sigma^2}\} \quad (1)$$

where parameter σ controls the size of neighborhood in the graph (discussed below).

Based on S, we can construct a fully connected, undirected graph whose vertices are the subjects and weights are the similarity between subjects pairs. Then, the normalized graph *Laplacian* [11] is computed as $L = T^{-0.5}(T - S)T^{-0.5}$, where T is diagonal matrix with element $t_{ii} = \sum_{j=1}^{83} s_{ij}$ for $i = 1, \ldots, 83$. The k (number of clusters) eigenvectors corresponding to the k smallest eigenvalues of L can be regarded as the feature vectors for all the subjects in the lower-dimensional space.

We set the number of clusters k equal to three based on Ding`s observation in his study [4]. By doing this, all the subjects are projected onto a sphere in $\mathbb{R}^3$. Subsequently, k-means clustering [13] is applied to divide subjects into three groups. Considering both brain hemispheres, six groups in total are generated. Since the k-means algorithm is randomly initialized, and may yield different partitions, we repeat k-means clustering 20 times and choose the partition with the highest average generalized DSC between the warped template segmentation and the subject`s segmentation (will be discussed in Sect. 3.3) to be the final partition.

3.2 Unbiased Population Averaging and Surface Mesh Generation

For each group, an unbiased population template is constructed from the multi-label segmentations by applying the iterative unbiased template building algorithm [14] and implementing the shape averaging approach in [16]. The metric optimized in this step during the subject-template registration is the same as the pairwise registrations above. Within each group, we choose the segmentation that is most similar to the others in its group (based on the pairwise similarity matrix) as the initial template to guide the template building process. The posterior probability maps for all the seven labels in the template space are used to vote to get the template segmentation.

For each group, a surface mesh is then generated for the union of the ERC, BA35 and BA36 labels, which are the structures of interest in this paper. As shown in Fig. 1 (d) and (e), the surface mesh is much smoother than the multi-atlas segmentation.

3.3 Thickness Measurement

Surface meshes are then warped back to the space of each subject using the corresponding diffeomorphic field computed in the template building step. Using this smooth surface approximation of the previous blocky labels (Fig. 1d,e), regional thickness can be computed by extracting the pruned Voronoi skeleton of the smooth mesh [17] and measuring the distance between each surface vertex and the closest point on the pruned Voronoi skeleton.

To measure how faithful the smooth template-based mesh approximation is to the input segmentations, we compute the average DSC between the multi-atlas segmentation of each subject and the segmentation obtained by warping the corresponding template's segmentation into the space of the subject.

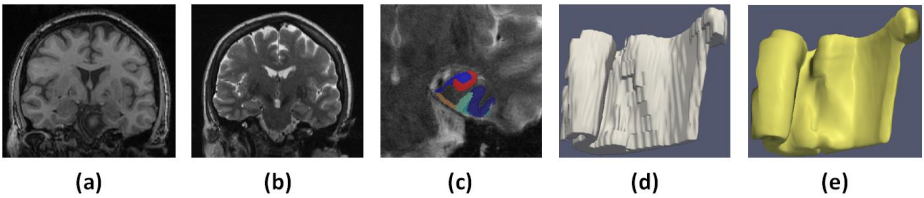


Fig. 1. Sagittal slice of T1-weighted image (a) and T2-weighted image (b). (c) Automatic segmentation. (d) Blocky automatic segmentation mesh. (e) Smooth surface mesh representation.

4 Experiments and Results

4.1 Volumetric and Thickness Measurements

We apply our method to the clinical dataset in Sect. 2 and measure the discriminative ability of the thickness measured obtained using the proposed "three-group" (TG) approach to three quantitative measures. As the first alternative, we measure thickness using a "single-group" (SG), which assigns all the subjects in each hemisphere to the same group, and thus does not account for PRC anatomical variation. As additional comparison measures, we (a) compute the normalized volume (volume of structure divided by its length of segmentation in the anterior-posterior direction) for ERC, BA35 and BA36 for both hemispheres directly from the multi-atlas segmentation and (b) extract a cortical thickness map from the T1-weighted MRI using an established method [9], and integrate this map over the ERC, BA35 and BA36 labels, which are first mapped into the space of the T1-weighted MRI using rigid registration.

4.2 Results

Among the 83 subjects, 30, 20 and 33 of them were clustered into group 1, 2 and 3 separately for the left hemisphere. On the right, the number of subjects in group 1, 2 and 3 are 26, 26 and 31 respectively. Fig. 2 shows the smooth meshes for all the six templates (three per hemisphere) from the TG approach. Group 1 templates resemble the continuous CS variant. Discontinuous CS is observed in group 2 and 3 which differ, as expected, by the relative length of the anterior and posterior CS (i.e. anterior CS is shorter in group 2 while it is longer in group 3). This indicates spectral clustering is able to automatically identify the three anatomical variants [7]. Fig. 2 also shows the meshes for SG, which look like a blend of the three meshes from TG. The odd shape of BA36 of the left SG template (indicated by the white circle in SG templates) is likely the result of ignoring the anatomical variation. The shapes of the templates indicate that a single template is limited in its ability to represent all subject segmentations well.

To evaluate this in a more quantitative way, we compute the average DSC for ERC, BA35, BA36 and CS between the warped template segmentation and subject`s segmentation in the space of each subject, which are shown in Table 1. As can be observed, the TG approach yields higher overlap for all the labels except ERC. Note

the dramatic increase in CS overlap, which indicates the warped meshes are better able to represent the segmentation using the TG approach. Another observation is the overlap remains almost the same for ERC. This demonstrates that TG approach does not degrade the measurement accuracy for a relatively consistent adjacent structure.

Table 1. Average for ERC, BA35, BA36 and CS between warped template segmentation and subject`s automatic segmentation using "single-group" (SG) and "three-groups" (TG) approaches (computed in subject space)

	Left Hemisphere				Right Hemisphere			
	ERC	**BA35**	**BA36**	**CS**	**ERC**	**BA35**	**BA36**	**CS**
SG	0.983 (± 0.005)	0.954 (± 0.015)	0.934 (± 0.022)	0.591 (± 0.151)	0.982 (± 0.010)	0.948 (± 0.014)	0.962 (± 0.016)	0.666 (± 0.123)
TG	0.983 (± 0.005)	0.965 (± 0.016)	0.959 (± 0.018)	0.803 (± 0.082)	0.983 (± 0.010)	0.958 (± 0.010)	0.970 (± 0.012)	0.749 (± 0.133)

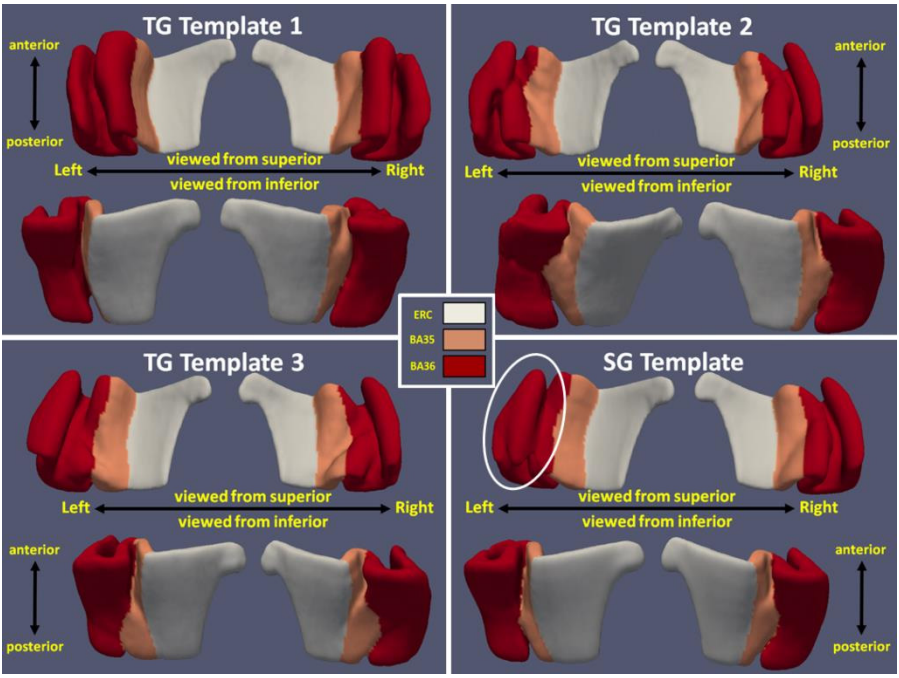


Fig. 2. Left and right surface meshes for SG and TG approaches viewing from superior and inferior. White circle shows the odd shape at the left SG template.

To further evaluate the proposed technique`s performance in clinical applications, we fit a general linear model to the thickness measurements with group membership, age, and intracranial volume as covariates, and report the t-statistic for the NC-MCI contrast. We also perform ROC analysis to the outputs of the four measurement approaches and report area under the curve (AUC) for group discrimination between MCI and NC groups. Intracranial volume is computed the same way as that in [6].

The thickness for each label is computed by integrating thickness value for all the vertices on the surface mesh belong to that label. The results are shown in Table 2.

From Table 2, it can be observed that SG and TG demonstrate stronger effects in distinguishing the two groups in ERC and BA35 (especially left BA35). However, thickness measurement using T1-weighted MRI turns out to be the best performer for BA36. Importantly, based on the work of Braak and Braak [2], as well as others, greater discrimination in ERC and BA35 is more biologically plausible given the earlier and greater neurofibrillary tangle pathology in these regions than BA36 [1]. Poorer performance in BA36 of SG and TG may result from poorer localization in the T1 approach. Overall, the results of SG and TG are more consistent with the known early pathology of this region in AD [1]. Comparing to SG, which shows relatively good performance, the TG approach, which accounts for anatomic variability, does appear to boost further the statistical power of thickness measurement. Another interesting observation is the left-right asymmetry in PRC, which shows up regardless of how we analyze the data (volumetry vs. thickness, T1 vs. T2) and might be explained by a bias towards verbal memory deficits in the MCI cohort.

Table 2. Statistical analysis results for the four measurements conducted in this paper

Label	Measurement	Left Hemisphere			Right Hemisphere		
		T-test	P-value	AUC	T-test	P-value	AUC
ERC	T1 Thickness	2.34	0.022	0.61	2.23	0.029	0.61
	Volume	2.64	0.0099	0.67	1.22	0.23	0.58
	SG Thickness	2.99	0.0037	0.66	2.75	0.0073	0.67
	TG Thickness	3.36	0.0012	0.68	2.73	0.0078	0.66
BA35	T1 Thickness	2.19	0.031	0.63	1.95	0.055	0.66
	Volume	4.46	2.6e-5	0.77	1.91	0.060	0.64
	SG Thickness	5.39	6.8e-7	0.82	2.31	0.023	0.67
	TG Thickness	5.58	3.1e-7	0.83	2.32	0.023	0.65
BA36	T1 Thickness	4.01	1.3e-4	0.73	1.44	0.15	0.60
	Volume	3.18	0.0021	0.68	-0.01	0.99	0.49
	SG Thickness	2.96	0.0040	0.67	0.70	0.49	0.55
	TG Thickness	3.32	0.0014	0.67	1.27	0.21	0.58

5 Conclusion

In this paper, we proposed a novel automatic clustering and thickness measurement pipeline for PRC based on automatic segmentation. For evaluation, we applied our technique to dataset of patients with MCI, often enriched in individuals with prodromal AD, and NC adults. The comparison between the surface meshes for TG and SG approaches demonstrates that group partitioning is a critical step to deal with anatomical variation within PRC, a key for accurately measuring thickness based on automatic segmentation. The statistical analysis supports the notion that the TG approach enhances power of discrimination between MCI and NC adults compared to volumetric measurement, the SG approach, and thickness measurement based on T1-weighted scans. As such, this method may have important utility in the early diagnosis and monitoring of AD, as well as providing accurate measurements to enhance brain-behavior studies of these regions.

References

1. Aggleton, J.P., Brown, M.: Interleaving brain systems for episodic and recognition memory. Trends CognSci. 10, 455–463 (2006)
2. Braak, H., Braak, E.: Staging of Alzheimer's disease-related neurofibrillary changes. Neurobiol Aging 16, 271–8; discussion 278–84 (1995)
3. Insausti, R., Juottonen, K., Soininen, H., Insausti, A.M., Partanen, K., Vainio, P., Laakso, M.P., Pitkänen, A.: MR volumetric analysis of the human entorhinal, perirhinal, and temporopolar cortices. AJNR Am. J. Neuroradiol. 19, 659–671 (1998)
4. Ding, S.L., Van Hoesen, G.W.: Borders, extent, and topography of human perirhinal cortex as revealed using multiple modern neuroanatomical and pathological markers. Human Brain Mapping 31(9), 1359–1379 (2010)
5. Augustinack, J.C., Huber, K.E., Stevens, A.A., Roy, M., Frosch, M.P., van der Kouwe, A.J.W., Wald, L.L., Van Leemput, K., McKee, A.C., Fischl, B.: Alzheimer's Disease Neuroimaging Initiative: Predicting the location of human perirhinal cortex, brodmann's area 35, from mri. Neuroimage 64, 32–42 (2013)
6. Yushkevich, P.A., Wang, H., Pluta, J., Das, S.R., Craige, C., Avants, B.B., Weiner, M.W., Mueller, S.: Nearly Automatic Segmentation of Hippocampal Subfields in In Vivo Focal T2-Weighted MRI. Neuroimage 53(4), 1208–1224 (2010)
7. Pluta, J., Yushkevich, P., Das, S., Wolk, D.: In vivo analysis of hippocampal subfield atrophy in mild cognitive impairment via semi-automatic segmentation of T2-weighted MRI. J. Alzheimers Dis. 29, 1–15 (2012)
8. Mueller, S.G., Weiner, M.W.: Selective effect of age, Apo e4, and Alzheimer's disease on hippocampal subfields. Hippocampus 19, 558–564 (2009)
9. Das, S.R., Avants, B.B., Grossman, M., Gee, J.C.: Registration based cortical thickness measurement. Neuroimage 45(3), 867–879 (2009)
10. Fischl, B.: Freesurfer. Neuroimage (2012)
11. Chung, F.R.K.: Spectral graph theory. Regional Conference Series in Mathematics, American Mathematical Society 92, 1–212 (1997)
12. Wolz, R., Aljabar, P., Hajnal, J.V., Hammers, A., Rueckert, D.: LEAP: learning embeddings for atlas propagation. NeuroImage 49(2), 1316–1325 (2010)
13. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press, Berkeley (1967)
14. Avants, B., Epstein, C., Grossman, M., Gee, J.: Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis 12, 26–41 (2008a)
15. Crum, W.R., Camara, O., Hill, D.L.G.: Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Trans. Med. Imaging 25, 1451–1461 (2006)
16. Joshi, S., Davis, B., Jomier, M., Gerig, G.: Unbiased diffeomorphic atlas construction for computational anatomy. Neuroimage 23(suppl. 1), S151–S160 (2004)
17. Ogniewicz, R.L., Kubler, O.: Hierarchic Voronoi skeletons. Pattern Recognit. 28, 343–359 (1995)

Constructing 4D Infant Cortical Surface Atlases Based on Dynamic Developmental Trajectories of the Cortex

Gang Li, Li Wang, Feng Shi, Weili Lin, and Dinggang Shen

Department of Radiology and BRIC, University of North Carolina at Chapel Hill, NC, USA

Abstract. Cortical surface atlases play an increasingly important role for analysis, visualization, and comparison of results across different neuroimaging studies. As the first two years of life is the most dynamic period of postnatal structural and functional development of the highly-folded cerebral cortex, longitudinal (4D) cortical surface atlases for the infant brains during this period is highly desired yet still lacking for early brain development studies. In this paper, we construct the *first* longitudinal (4D) cortical surface atlases for the dynamic developing infant cortical structures at 1, 3, 6, 9, 12, 18 and 24 months of age, based on 202 serial MRI scans from 35 healthy infants. To ensure longitudinal consistency and unbiasedness of the 4D infant cortical surface atlases, we first compute the within-subject mean cortical folding geometries by groupwise registration of longitudinal surfaces of each infant. Then we establish inter-subject cortical correspondences by groupwise registration of the within-subject mean cortical folding geometries of all infants. More importantly, for the first time, we further parcellate the 4D infant surface atlases into developmentally and functionally distinctive regions based solely on the dynamic developmental trajectories of the cortical thickness, by using the spectral clustering method. Specifically, to deal with the problem that each infant has different number of scans, we first compute the within-subject affinity matrix of vertices' cortical thickness trajectories of each infant, and then we use the averaged affinity matrix of all infants for parcellation. Our constructed 4D infant cortical surface atlases with developmental trajectories based parcellation will greatly facilitate the surface-based analysis of dynamic brain development in infants.

Keywords: Infant, cortical surface, atlas construction, surface parcellation.

1 Introduction

A brain atlas is a representation of anatomical structures and other reference information in a spatial framework, providing a useful repository of knowledge and also facilitating the analysis of spatially-localized experimental data [1]. In neuroimaging studies, brain atlases play an increasingly important role for analysis, visualization, and comparison of results across different studies [1]. Cortical surface-based analysis is particularly suitable for studying the highly folded cerebral cortex, as this type of methods respects the topology of the cortex and facilitates registration, analysis, and visualization of buried sulci [1, 2]. Moreover, cortical surface based measurements, e.g., cortical thickness, surface area, and gyrification, provide very detailed aspects of

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 89–96, 2014.

the cortex. Accordingly, several cortical surface atlases have been developed [1-3]. For example, FreeSurfer surface atlas is built by landmark-free registration of spherical surfaces of 40 adult brains based on the cortical folding geometries [2]. PALS-B12 and PALS-term12 surface atlases are built by using the sulcal-gyral landmarks constrained registration of 12 adult brains and 12 neonatal brains [3], respectively.

The first two years of life is an exceptionally dynamic period for the structural and functional development of the cerebral cortex [4]. For cortical surface-based analysis of early brain development [3-6], longitudinal (4D) cortical surface atlases for characterizing the dynamic developing infant cortical structures are highly desired yet still lacking. Meanwhile, the existing parcellation in cortical surface atlases is typically defined based on the sulcal-gyral landmarks [7], which, however, are problematic for reliable and precise localization of functional regions. This is because sulcal-gyral patterns are extremely variable and poorly match with the microstructurally defined borders [8]. The microstructures, reflecting the molecular organization and functional principles of the cortex, and their connectivity jointly determine the functional role of a cortex region [8]. Therefore, the microstructurally derived surface atlas parcellation is more ideally for studying the cortex. For example, the well-known Brodmann's map and the recent JuBrain atlas are all defined based on cytoarchitecture [8]. Though the current MRI cannot reveal the cytoarchitecture, parcellation of the surface atlas based on the genetic correlations of cortical thickness has been developed using the MRI of adult twins, reflecting the genetic influences on cortical regionalization [9].

Motivated by these works, in this paper, we construct the *first* longitudinal (4D) cortical surface atlases to characterize the dynamic developing cortical structures at 1, 3, 6, 9, 12, 18 and 24 months of age, based on 202 serial MRI scans from 35 healthy infants. Meanwhile, the dynamic cortical developmental trajectories in infants reflect the underlying cytoarchitectonic changes of the cortex (e.g., increases in dendritic arborization, axonal elongation and thickening, synaptogenesis and glial proliferation [5]), and thus can help better define the microstructurally distinctive cortical regions than using sulcal-gyral landmarks. Therefore, for the first time, we further parcellate the 4D cortical surface atlases into developmentally and functionally distinctive regions based solely on the developmental trajectories of cortical thickness in infants.

2 Methods

2.1 Data Acquisition and Image Processing

Serial T1-, T2-, and diffusion-weighted MR images of 35 healthy infants (18 males/17 females) were acquired using a Siemens 3T MR scanner. Each infant was scheduled to be scanned at every 3 months in the first year and then every 6 months in the second year. Due to insufficient-quality and uncompleted scans, each infant has different number of scans, ranging from 4 to 7 in the first two years. In total, 202 quality scans from 35 infants were acquired, with each infant having 5.8 scans on average. The numbers of scans were 35 at 1 month, 28 at 3 months, 31 at 6 months, 27 at 9 months, 29 at 12 months, 31 at 18 months, and 21 at 24 months, respectively.

T1 images (144 sagittal slices) were acquired with the imaging parameters: TR/TE = 1900/4.38 ms, flip angle = 7, resolution = $1 \times 1 \times 1$ mm^3. T2 images (64 axial

slices) were acquired with the parameters: TR/TE = 7380/119 ms, flip angle = 150, resolution = $1.25 \times 1.25 \times 1.95$ mm^3. Diffusion-weighted images (DWI) (60 axial slices) were acquired with the parameters: TR/TE = 7680/82 ms, resolution = $2 \times 2 \times 2$ mm^3, 42 non-collinear diffusion gradients, and diffusion weighting b =1000s/mm^2.

T2 images and fractional anisotropy (FA) images, derived from distortion corrected DWI, were rigidly aligned onto their T1 images and further resampled to $1 \times 1 \times 1$ mm^3. After removing non-cerebral tissues, tissue segmentation was performed using a 4D level-set method, which integrates the multimodal information of T1, T2 and FA images [10]. Non-cortical structures were filled, and each brain was separated into left and right hemispheres. For each hemisphere, inner and outer cortical surfaces were reconstructed with the following steps: (1) performing topology correction of white matter (WM); (2) tessellating the boundary of the corrected WM as an explicit triangular mesh representation; (3) applying a deformable surface method to obtain the refined inner and outer surfaces [4]. The cortical thickness of each vertex was computed as the mean of the minimum distance from inner to outer surfaces and that from outer to inner surfaces. The inner surface was further mapped to a standard sphere [2].

2.2 Constructing 4D Infant Cortical Surface Atlases

One solution for 4D cortical surface atlases construction is to independently align the spherical surfaces of all infants at each age with groupwise registration. However, this ignores the within-subject longitudinal constraints and thus could lead to temporally inconsistent atlases. To ensure longitudinal consistency and unbiasedness of 4D infant cortical surface atlases, the proposed method makes use of the within-subject longitudinal constraints and consists of the following 5 steps, as shown in Fig. 1. **First**, for each hemisphere, all longitudinal spherical surfaces of the same infant were groupwisely aligned using Spherical Demons [11], to establish the unbiased within-subject cortical correspondences (step 1 in Fig. 1). **Second**, for each infant, the within-subject mean cortical folding geometries were computed based on their cortical correspondences. As major cortical folding were present at term birth and were preserved during postnatal development [3, 4], the within-subject mean cortical folding geometries are sharp and contain detailed information of cortical folding. **Third**, the within-subject mean cortical folding geometries of all infants were groupwisely aligned using Spherical Demons, to establish unbiased inter-subject cortical correspondences (step 3 in Fig. 1). **Fourth**, for each age, the inter-subject cortical correspondences were established based on the correspondences defined by their within-subject mean cortical folding geometries, and each surface was resampled to a standard-mesh tessellation, leading to the consistent inter-subject cortical correspondences across all ages. **Finally**, for each age, a surface atlas consisting of the mean and variance of cortical folding geometries, e.g., mean curvature, average convexity, and mean curvature of inflated surfaces [2], across all infants at this age was constructed on the spherical surface (step 5 in Fig. 1). Thus, the 4D infant surface atlases capture both the longitudinally-consistent changes and inter-subject variability of cortical folding.

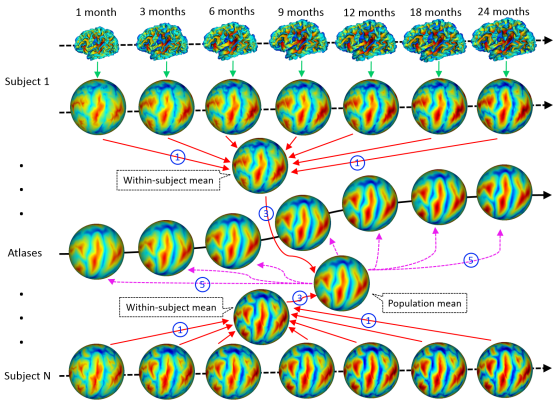


Fig. 1. Illustration of the proposed method for 4D infant cortical surface atlas construction

2.3 Parcellation of 4D Surface Atlases Using Developmental Trajectories

As mentioned, the dynamic cortical developmental trajectories in infants reflect the underlying microstructural changes of the cortex [4], and thus could help better define the microstructurally and functionally distinctive regions than using sulcal-gyral landmarks. Accordingly, we proposed to parcellate 4D surface atlases based solely on the dynamic developmental trajectories of cortical thickness, which is well correlated with cognitive functions. Before performing parcellation, for each time point of each infant, the cortical thickness map was smoothed and normalized by the mean cortical thickness of the cortical surface to account for dynamic cortex changes in infants.

We adopted the spectral clustering method [12] for 4D surface atlases parcellation. First, for each infant, we defined the *subject-specific* affinity matrix of the developmental trajectories of cortical thickness between each pair of vertices using Pearson's correlation, to address the problem that each infant had different number of scans and regional variations of cortical thickness. Specifically, for an infant subject s with n_s time points, at a vertex i, its cortical thickness trajectory was denoted as $X_{i,t}^s$, $t \in \{1, \dots, n_s\}$. For two vertices i and j, Pearson's correlation coefficient was computed as: $r^s(i,j) = \frac{\sum_{t=1}^{n_s}(X_{i,t}^s - \overline{X_i^s})(X_{j,t}^s - \overline{X_j^s})}{\sqrt{\sum_{t=1}^{n_s}(X_{i,t}^s - \overline{X_i^s})^2}\sqrt{\sum_{t=1}^{n_s}(X_{j,t}^s - \overline{X_j^s})^2}}$. Thus, for each infant, an affinity matrix A^s can be computed, with its element $A_{i,j}^s = 1 + r^s(i,j)$. The range of $A_{i,j}^s$ is between 0 and 2, indicating that two vertices are very different or very similar. Note that $A_{i,j}^s$ was defined for each pair of vertices and relied solely on the developmental trajectories.

Then, we computed the mean affinity matrix A by averaging the corresponding elements of affinity matrices of all infants. Next, we normalized the affinity matrix A as $L = D^{-1/2}AD^{-1/2}$, where D was a diagonal matrix with $D_{i,i} = \sum_j A_{i,j}$. The data was then represented in an eigenspace using top 30 eigenvectors of D, which better captured the distributions of the original data points. Finally, we used Gaussian mixture models, which was initialized by k-means method, to cluster the data into different groups using the new data representation in the eigenspace.

3 Results

We utilize 35 healthy infants with 202 MRI scans in total, for construction of 4D infant cortical surface atlases. Fig. 2 shows means and variations of the average convexity and mean curvature on the spherical spaces of the constructed 4D infant cortical surface atlases for the left hemisphere at 1, 3, 6, 9, 12, 18 and 24 months of age. The mean curvature reflects the fine-scale geometry of the cortical folding, while the average convexity, which records the accumulated movement for each vertex during surface inflation, reflects the large-scale geometry of the cortical folding [2]. As we can see, major cortical folding patterns are temporally quite consistent for both average convexity and mean curvature in the first 24 months. However, the magnitude of the average convexity increases considerably, though the mean curvature changes relatively subtly. The central sulcus consistently exhibits small inter-subject variations of cortical folding geometries, while the prefrontal and parietal cortices show large inter-subject variations.

To demonstrate the advantage of the proposed method for 4D infant cortical surface atlas construction, we compare it with the conventional case of constructing the 4D surface atlases by independent groupwise registration of all surfaces at each time point. Fig. 3 provides a representative example of the close-up view of the mean curvatures on 4D cortical surface atlases constructed by two different methods. As we can see, the proposed method generates longitudinally more consistent 4D cortical surface atlases, along with much shaper cortical folding patterns, than the counterpart method that uses independent groupwise registration of surfaces for each time point.

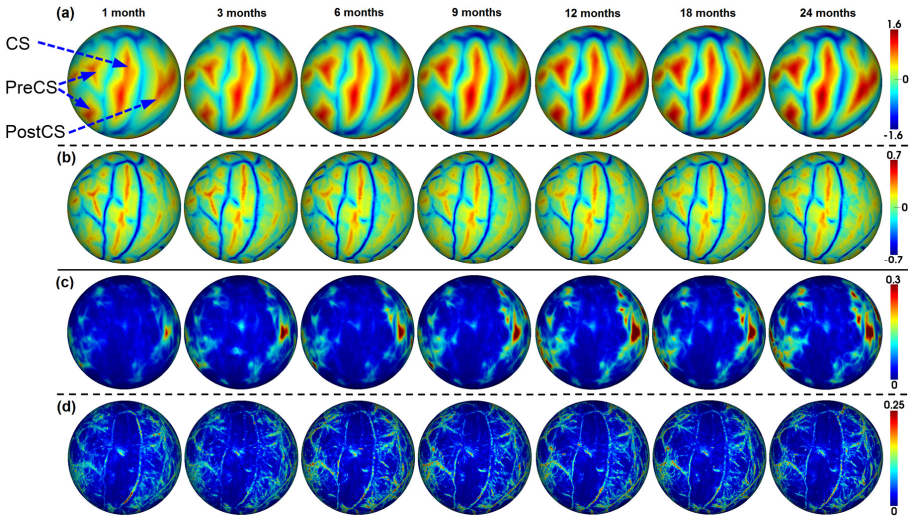


Fig. 2. Longitudinal (4D) infant cortical surface atlases at 1, 3, 6, 9, 12, 18 and 24 months of age. (a) Average convexity. (b) Mean curvature. Blue color indicates gyri, and red color indicates sulci. CS=central sulcus, PreCS=precentral sulcus, PostCS=postcentral sulcus. (c) Variations of average convexity. (d) Variations of mean curvature.

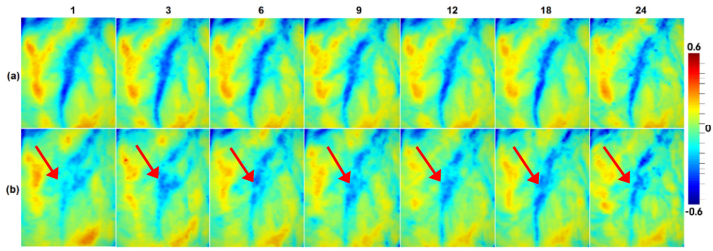


Fig. 3. Comparisons of 4D surface atlases (color-coded by the mean curvature), constructed by (a) the proposed method and (b) groupwise surface registration at each age independently

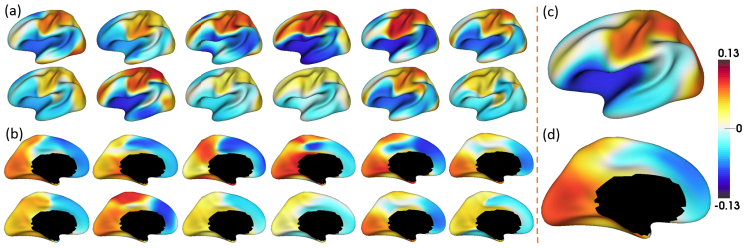


Fig. 4. Average Pearson's correlation map of the developmental trajectories of the cortical thickness between each vertex and all other vertices. (a)-(b) Correlation maps of 12 randomly selected infants. (c)-(d) Mean correlation maps of 35 infants.

Fig. 4 shows the average Pearson's correlation maps of the developmental trajectories of cortical thickness between each vertex and all other vertices of 12 randomly selected infants, along with the mean correlation maps of all 35 infants. Individual variations of correlation patterns may come from inherent inter-subject variability or different number of scans. However, the global correlation patterns (positive vs negative correlations) of individuals are quite similar with those of the averaged patterns of all infants. The motor, somatosensory, and visual cortices have strong positive correlations (large similarity) with other regions, while temporal and prefrontal association cortices have strong negative correlations (small similarity) with other regions in their trajectories. Of note, the mean correlation map may potentially be confounded by regional variations of surface alignment errors, e.g., association cortices with more variable folding may have lower surface alignment accuracies than unimodal cortices.

Fig. 5 shows the parcellation of the surface atlas of the left hemisphere with 2 and 12 clusters, respectively, based on the developmental trajectories of cortical thickness, in order to compare with the recent surface atlas parcellation based on the genetic correlation of cortical thickness derived from MRI of adult twins [9]. In the 2-cluster parcellation, the motor, somatosensory and visual cortices are grouped into a cluster, while the prefrontal and temporal association cortices are grouped into the other cluster, indicating distinct developmental trajectories between high-order association cortices and unimodal cortices. In the 12-cluster parcellation, all clusters correspond closely to the functionally meaningful regions, including: 1, sensorimotor cortex; 2, superior parietal cortex; 3, inferior parietal cortex; 4, perisylvian region; 5, occipital cortex; 6, cingulate cortex; 7, orbitofrontal cortex; 8, insula cortex; 9, medial temporal

cortex; 10, lateral temporal cortex; 11, dorsolateral prefrontal cortex; 12, medial frontal cortex. Notably, boundaries of many clusters do not follow sulcal-gyral landmarks, suggesting that they may not be borders for the microstructural development. To further confirm the surface atlas parcellation, Fig. 6 show seeds (25 different seeds) based analysis of mean correlation maps of the cortical thickness trajectories. At each seed location on the large (grey) cortical surface, its correlations with all other vertices are shown by a respective color-coded small surface map. High correlations are always observed in the region closest to each seed, while the seeds in the same cluster yield largely the similar correlation patterns. Meanwhile, the boundaries of positive and negative correlations are largely similar, regardless of the seed position. In addition, as shown in Fig. 5, several clusters in our parcellation and the parcellation based on genetic correlation of adult cortical thickness [9] look similar, e.g., perisylvian region (cluster 4), dorsolateral prefrontal cortex (cluster 11) and medial temporal cortex (cluster 9), which, however, needs further quantitative comparison.

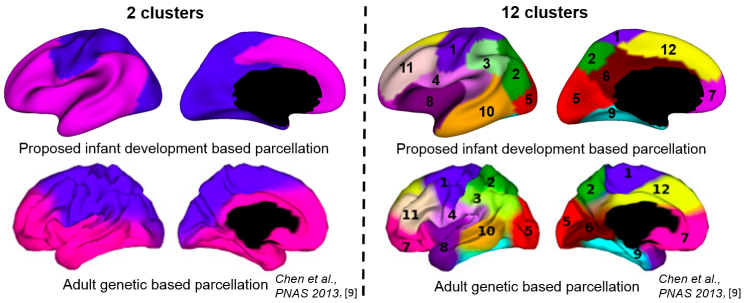


Fig. 5. Cortical surface atlas parcellations based on infant trajectories of cortical thickness, compared with the parcellations based on adult genetic correlation of cortical thickness [9]

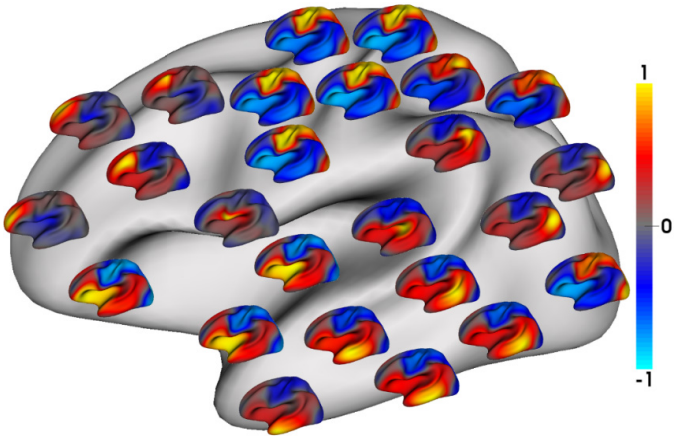


Fig. 6. Seed-based analysis of group averaged correlation patterns of cortical thickness trajectories from 35 infants, for confirming the parcellation in Fig. 5. For each of 25 seeds, its correlations with all other vertices are shown as a respective color-coded small surface map.

4 Conclusion

Our contribution in this paper is threefold. *First*, we proposed a new method for building longitudinally-consistent 4D cortical surface atlases. *Second,* we constructed the first 4D infant cortical surface atlases for characterizing the dynamic developing cortex at 1, 3, 6, 9, 12, 18 and 24 months of age. *Third*, for the first time, we parcellated the 4D infant surface atlases into developmentally and functionally distinctive regions based solely on the dynamic trajectories of the cortical thickness. In our future work, we will also consider using the developmental trajectories of the local surface area and local gyrification for parcellation. We will make our 4D infant cortical surface atlases publically available for facilitating early brain development studies.

References

1. Van Essen, D.C., Dierker, D.L.: Surface-based and probabilistic atlases of primate cerebral cortex. Neuron 56, 209–225 (2007)
2. Fischl, B., Sereno, M.I., Tootell, R.H., Dale, A.M.: High-resolution intersubject averaging and a coordinate system for the cortical surface. Human Brain Mapping 8, 272–284 (1999)
3. Hill, J., Dierker, D., Neil, J., Inder, T., Knutsen, A., Harwell, J., Coalson, T., Van Essen, D.: A surface-based analysis of hemispheric asymmetries and folding of cerebral cortex in term-born human infants. J. Neurosci. 30, 2268–2276 (2010)
4. Li, G., Nie, J., Wang, L., Shi, F., Lin, W., Gilmore, J.H., Shen, D.: Mapping region-specific longitudinal cortical surface expansion from birth to 2 years of age. Cereb. Cortex 23, 2724–2733 (2013)
5. Rodriguez-Carranza, C.E., Mukherjee, P., Vigneron, D., Barkovich, J., Studholme, C.: A framework for in vivo quantification of regional brain folding in premature neonates. Neuroimage 41, 462–478 (2008)
6. Xue, H., Srinivasan, L., Jiang, S., Rutherford, M., Edwards, A.D., Rueckert, D., Hajnal, J.V.: Automatic segmentation and reconstruction of the cortex from neonatal MRI. Neuroimage 38, 461–477 (2007)
7. Desikan, R.S., Segonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J.: An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968–980 (2006)
8. Zilles, K., Amunts, K.: TIMELINE Centenary of Brodmann's map - conception and fate. Nature Reviews Neuroscience 11, 139–145 (2010)
9. Chen, C.H., Fiecas, M., Gutierrez, E.D., Panizzon, M.S., Eyler, L.T., Vuoksimaa, E., Thompson, W.K., Fennema-Notestine, C., Hagler Jr., D.J., Jernigan, T.L., Neale, M.C., Franz, C.E., Lyons, M.J., Fischl, B., Tsuang, M.T., Dale, A.M., Kremen, W.S.: Genetic topography of brain morphology. Proc. Natl. Acad. Sci. U S A 110, 17089–17094 (2013)
10. Wang, L., Shi, F., Yap, P.T., Gilmore, J.H., Lin, W., Shen, D.: 4D multi-modality tissue segmentation of serial infant images. PLoS One 7, e44596 (2012)
11. Yeo, B.T., Sabuncu, M.R., Vercauteren, T., Ayache, N., Fischl, B., Golland, P.: Spherical demons: fast diffeomorphic landmark-free surface registration. IEEE Trans. Med. Imaging 29, 650–668 (2010)
12. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. Adv. Neur. In. 14, 849–856 (2002)

Low-Rank to the Rescue – Atlas-Based Analyses in the Presence of Pathologies

Xiaoxiao Liu[1], Marc Niethammer[2], Roland Kwitt[3], Matthew McCormick[1], and Stephen Aylward[1]

[1] Kitware Inc., USA
[2] University of North Carolina at Chapel Hill, USA
[3] Department of Computer Science, University of Salzburg, Austria

Abstract. Low-rank image decomposition has the potential to address a broad range of challenges that routinely occur in clinical practice. Its novelty and utility in the context of atlas-based analysis stems from its ability to handle images containing large pathologies and large deformations. Potential applications include atlas-based tissue segmentation and unbiased atlas building from data containing pathologies. In this paper we present atlas-based tissue segmentation of MRI from patients with large pathologies. Specifically, a healthy brain atlas is registered with the low-rank components from the input MRIs, the low-rank components are then re-computed based on those registrations, and the process is then iteratively repeated. Preliminary evaluations are conducted using the brain tumor segmentation challenge data (BRATS '12).

1 Introduction

Image-based lesion detection and segmentation are needed to assess and plan the treatment of patients suffering from traumatic brain injuries (TBI), brain tumors, or stroke [2]. One popular method for such image analysis involves registering an atlas to the patient's images to estimate tissue priors. However, if the patient's images contain large pathologies, then lesion-induced deformations may inhibit atlas registration and confound the tissue priors. Furthermore, even forming an appropriate, unbiased atlas for segmentation may be problematic.

In unbiased atlas building [3], when images with lesions are used to form the atlas, the lesions propagate into and corrupt the atlas. However, particularly for research projects with limited time and financial resources or involving a new imaging protocol or children, it can be problematic to obtain a sufficient number of protocol-matched scans from healthy subjects for atlas formation. Hence, registration methods tolerant to such image corruptions are desirable.

The iterative, low-rank image registration framework presented in this paper tolerates the presence of large lesions during image registration, and it can thereby aid in atlas-based segmentation and unbiased atlas formation by mitigating the effects described above. While our approach is general, in this paper we focus on registration in the presence of pathologies for the purpose of atlas-based tissue segmentation for illustration.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 97–104, 2014.

The most straightforward method to eliminate a lesion's influence during registration is to "mask" it so that the lesion's voxels are not considered during the computation of the image similarity metric. Other methods attempt to address this problem by joint registration and segmentation which tolerates missing correspondences [1], geometric metamorphosis that separates estimating healthy tissue deformation from modeling tumor change [5], or personalized atlas construction that accounts for diffeomorphic and non-diffeomorphic changes [9]. While effective, these methods require explicit lesion segmentations or initial lesion localizations, which, in this case, is actually the goal of the process.

Contribution. We propose to exploit *population information* to assess which parts of an image are likely lesions (they are inconsistent with the population) and which parts of an image should be considered *normal*. We adopt a recent machine learning technique, i.e., the decomposition of matrices into a low-rank and sparse components [6], in an iterative registration process to achieve this objective.

2 Low-Rank Plus Sparse Decomposition

In [6], Peng et al. propose to decompose a matrix of vectorized images into the sum of a low-rank and a sparse component (containing residuals) in the context of simultaneous image alignment. The intuition is that the portion of each image that cannot be explained by the low-rank model is allocated to the sparse part. Hence, the low-rank component could be interpreted as a blending of recorded values and values inferred from the population; the sparse component then contains each subject's *anomalous* values. Technically, the allocation of image intensities to each of those components is driven by the amount of linear-correlation across the images. Given a collection of n images having m voxels, we have:

D a $m \times n$ matrix in which each image I_i is a column vector that contains the m spatially-ordered voxel intensities in I_i.
L a $m \times n$ matrix that contains the low-rank representations L_i for each of the images in the collection D.
S a $m \times n$ matrix that is the sparse component, s.t. $S_i = D_i - L_i$.

The low-rank representation of D is then defined as

$$\{L^*, S^*\} = \arg\min_{L,S}(\|L\|_* + \lambda\|S\|_1) \text{ s.t. } D = L + S \ , \tag{1}$$

where $\|L\|_*$ is the nuclear norm of L and $\|S\|_1$ is the 1-norm of S. Since the problem is convex, a globally optimal solution $\{L^*, S^*\}$ can be obtained using, e.g., an augmented Lagrangian approach [4].

3 Integrate Low-Rank Decomposition into an Iterative Registration Framework

We have integrated the low-rank plus sparse decomposition into an iterative registration framework in which a group of input images, potentially containing

large pathologies and deformations, are registered to a normal-control atlas. Our premise is that by identifying the low-rank and sparse components of each input image, its low-rank component, which contains reduced or eliminated pathologies, can be more accurately registered with a normal-control atlas, compared to the *direct* registration of an image containing a pathology to an atlas.

The low-rank plus sparse decomposition exploits the fact that lesions generally do not manifest in consistent locations or with consistent appearance in populations. These inconsistencies result in lesions being reduced in the low-rank component and allocated to the sparse component. Thereby, the sparse component can be used to inform spatial and intensity priors for localizing and segmenting lesions.

Our method also supports unbiased atlas formation using data containing pathologies. Specifically, in the above framework the normal-control atlas I_A can be replaced by the mean low-rank image, at each iteration. In unbiased atlas-building the goal is to estimate an atlas image such that it is central with respect to the data population. This is achieved by minimizing

$$E(\{\Phi_i^{-1}\}, I_A) = \sum_{i=1,\ldots,N} Reg[\Phi_i^{-1}] + \sigma^{-2} Sim[I_i \circ \Phi_i^{-1}, I_A], \tag{2}$$

with respect to the unknown atlas image I_A and the unknown transformations $\{\Phi^{-1}\}$. Here, $Reg[\Phi]$ denotes a regularity measure for the transformation Φ, typically penalizing spatially non-smooth transformations and $Sim[I, J]$ is a chosen similarity measure between the images I and J. This could simply be the sum-of-squared intensity differences (SSD).

To optimize this energy using alternating optimization, we first keep I_A fixed while solving for $\{\Phi_i^{-1}\}$ and subsequently keep the transformations $\{\Phi_i^{-1}\}$ fixed while solving for I_A. The first part performs independent pairwise registrations between $\{I_i\}$ and the fixed image I_A. The second part requires, for SSD, minimizing

$$E(I_A) = \sum_{i=1,\ldots,N} \|I_i - I_A\|^2 \tag{3}$$

which is achieved by the mean image $I_A = {}^1\!/_N \sum_i I_i$. However, when lesions are present in $\{I_i\}$, the mean image is degraded by the undesired involvement of the lesions in the average. We can instead minimize Eq. (1) to obtain the low-rank approximations $\{L_i\}$ of the warped images $\{D_i = I_i \circ \Phi^{-1}\}$, and then minimize

$$E(I_A) = \sum_{i=1,\ldots,N} Sim[L_i, I_A] \tag{4}$$

with respect to the unknown atlas I_A. Again, for SSD, the solution will be the mean over $\{L_i\}$[1]. Other similarity measures, such as normalized cross correlation, could be used for registration, but they may require more challenging optimizations and may not be meaningful for atlas construction as they may

[1] In case of SSD it is advisable to initially histogram-normalize each I_i.

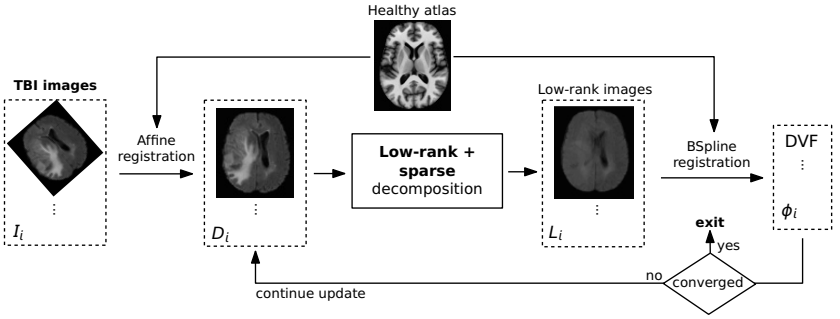


Fig. 1. An illustration of the proposed low-rank iterative image registration framework, where I_i refers to the i-th input image, D_i is the i-th vector of the input matrix D, L_i is low-rank component of the i-th input image and ϕ_i refers to the i-th registration map generated from BSpline image registration at each iteration.

make the atlas-image non-unique. Also note that when fixing the atlas, I_A, unbiased atlas construction simplifies to group-wise registration. The group-wise approach is essential because it allows for the population-based decomposition of the images into low-rank/sparse components (cf. §2). A general framework for our method is shown in Fig. 1. The algorithm proceeds as follows:

(1) Solve for affine transform $(\phi_i^0)^{-1}$ registering each I_i to the atlas image I_A.
(2) For each iteration j, compute the low-rank image L_i^j by solving Eq. (1).
(3) Solve for deformable transform $(\phi_i^j)^{-1}$ registering low-rank images L_i^j to I_A.
(4) Compose and apply transforms to I_i, s.t. $I_i^{j+1} = I_i \circ (\phi_i^0)^{-1} \cdots \circ (\phi_i^j)^{-1}$.
(5) Set $j \leftarrow j + 1$ and continue with step (2) until convergence.

Given a low-rank plus sparse decomposition, the registration step can be based on any standard deformable registration algorithm and its associated convergence characteristics apply. In our experiments, BSpline transforms and the Mattes mutual information (MMI) metric are used to register the low-rank images with the atlas, cf. step (3). The number of BSpline control points is increased gradually over the iterations to effect a coarse-to-fine optimization strategy. At each iteration we are maximizing the mutual information between the atlas image and each individual low-rank image, cf. Eq. (4). As our algorithm alternates between low-rank decomposition and registration, it can be considered a greedy strategy. Convergence is reached when the total change in deformation is small. In our experiments with the BRATS '12 dataset (8 inputs), results converge within 10 iterations. For the TumorSim data [7] (20 simulated T1 images), λ in Eq. (1) is set to 0.5 and for the patient data (8 FLAIR images), λ is set to 0.8.

4 Experimental Study

We have conducted initial assessments of our method for atlas-to-image registration using two evaluation metrics with simulated and patient data.

Quantitative Assessment of Atlas-to-Image Registration: The premise of atlas-based segmentation is that by registering an atlas with a target image, the tissue labels in the atlas provide spatial priors for the tissues in the target image. When atlas-to-image registration is successful, the atlas' tissue labels should align with the corresponding tissues in the target image. Therefore we compute the standard deviation of the target image intensities under each tissue label in the atlas. Smaller *tissue-class standard deviation (TCSD)* values indicate more accurate atlas-to-image registrations.

We calculate TCDS at each iteration to evaluate the convergence of the iterative framework. We also use it to compare our method with traditional BSpline atlas-to-image registration.

Qualitative Assessment for Lesion Segmentation: The iterative atlas-to-image registration process can be examined by inspecting the parts of each image allocated to its sparse component in each iteration. This *sparse image* at the final iteration (after reaching convergence) should be sensitive and specific to the lesion. By reviewing the sparse image's evolution over the iterations, we can qualitatively assess the effectiveness of our method in matching each patient's image to the healthy atlas, while not burdened by lesions.

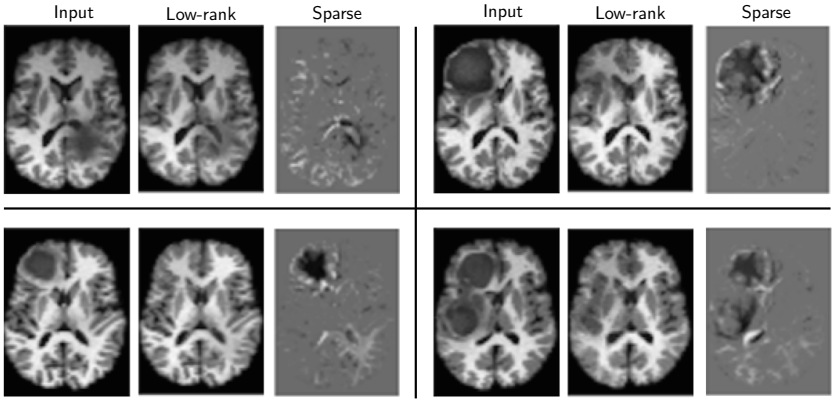
Note that only a qualitative assessment of lesions is made. Sparse images will contain some normal anatomic variation as well as the lesions. Over the iterations, the variations between the individual patients and the healthy atlas are minimized via the deformable registrations between the low-rank images and the healthy atlas. The sparse images after convergence could then serve as a strong prior for subsequent tumor segmentation algorithms, but lesion's heterogeneity as well as "normal" small-scale anatomic variations must be appropriately handled by subsequent lesion segmentation algorithms.

4.1 Case Studies and Results

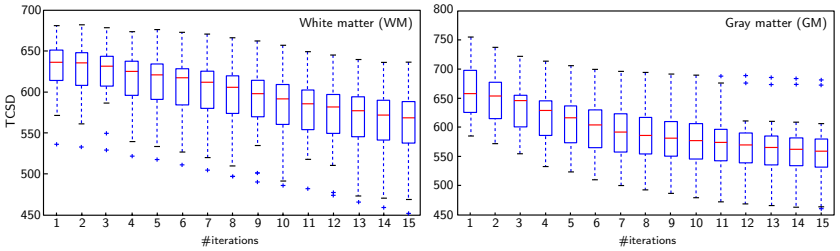
For the following case studies we used the SRI24 atlas [8] as the healthy atlas for registration and to provide gray-matter (GM), white-matter (WM), and cerebrospinal fluid (CSF) tissue labels after registration to compute the TCSD metric.

Case 1 (simulation data): The training data in the BRATS '12 challenge included MRI scans into which simulated high-grade and low-grade glioma tumors were injected using *TumorSim* [7]. We selected 20 cases containing large tumors and large deformations to form a set of challenging image-to-atlas registration tasks. Fig. 2(a) shows the first 4 subjects and their corresponding low-rank and sparse components during the 1^{st} iteration. Fig. 2(b) shows the TCSD values for the GM and WM classes after each iteration. The tightening of the statistics in Fig. 2(b) illustrates the convergence of the registration between each subject and the healthy atlas over time.

Case 2 (clinical data): A subset of 8 FLAIR images from BRATS '12 challenge are tested using the same experimental setup as the simulated data. Fig. 3(a)



(a) 4/20 simulated T1 brain volumes with the corresponding low-rank and sparse components (during the 1st iteration).

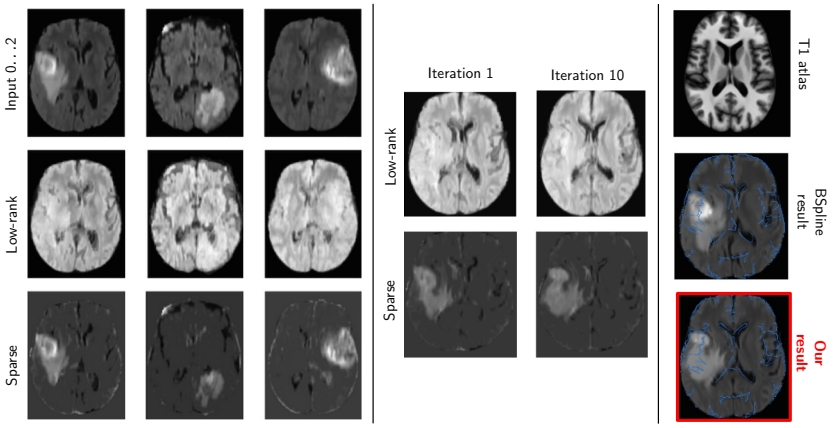


(b) Change in TCSD for WM and GM labels transcribed from the atlas.

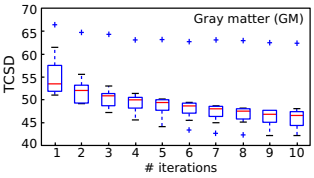
Fig. 2. Results on TumorSim dataset from BRATS '12

illustrates the input data and the results. We show the low-rank and sparse components of the first three input FLAIR images at the initial iteration in Fig. 3(a) (left). The low-rank plus sparse decomposition results at the 2nd and the 10th (final) iteration of the first input patient are shown in Fig. 3(a) (middle). On the right-hand side of Fig. 3(a), the top image is the normal-control SRI24 atlas T1 image that we used as the fixed normal-control image during each registration. As we can see from the box plots in Fig. 3(b) , the overall TCSD for the GM class improves over the iterations (each box contains the TCDS values of all 8 patients at each iteration).

To compare with a direct BSpline registration, we used the same BSpline parameters, e.g., number of control points, that are used in the final iteration of our method. We exclude the tumor region when calculating TCDS for each tissue class. The middle image of Fig. 3(a) is the atlas CSF label contour overlaid on the direct BSpline-registered (i.e., deform the original input FLAIR image to match the healthy atlas) image; the bottom image is the atlas CSF label contour overlaid on top of the converged deformed input image. Much better alignments are seen on the edges of the ventricles than it is the case for the middle image. The direct BSpline registration produces larger TCSD values for the GM class



(a) *Left*: first three FLAIR images and their initial low-rank and sparse components. *Middle*: low-rank and sparse components at the 2^{nd} and 10^{th} iteration of the first input. *Right*: SRI24 T1 atlas (top), atlas CSF label contour overlaid on the direct BSpline registered image (middle), atlas CSF label contour overlaid on the deformed input image at final iteration using our approach (bottom).



Patient	1	2	3	4	5	6	7	8
Ours	46.7	42.2	44.9	48.0	46.3	43.0	47.3	62.3
BSpline	68.0	48.3	50.8	53.4	40.5	53.7	58.2	39.4

(b) *Left*: Change in the TCSD for GM labels transcribed from the atlas after each iteration of our method. *Right*: TCSD comparison to traditional BSpline atlas-to-image registration for the GM class.

Fig. 3. Results on TBI patient dataset from BRATS '12

in most cases in this study, as shown in the table of Fig. 3(b). Our method performs worse on two cases (patient 5 and 8). Different from others they both have much narrower and distorted ventricles, which are high-contrast landmarks for guiding the registration optimization. Due to their distinctive appearance, the decomposed low-rank images contains very little truth geometries (mostly assigned to the sparse images), therefore the registration based on the low-rank component is not reliable. If more similar type datasets are included, the decomposition would be more effective and registration would have been resolved. Furthermore, only eight cases is not sufficient to represent a population, further work is needed to determine if and how many additional cases would be needed to represent a healthy population given subjects with pathologies.

5 Discussion and Future Work

The novel contributions of this paper are 1) the integrated formulation of low-rank image decomposition into atlas formation, 2) the use of low-rank image decomposition in atlas-to-image registration, and 3) the use of low-rank image decomposition as a prior for lesion identification and segmentation. These contributions are significant, because they allow images containing pathologies to drive atlas formation and they allow images containing pathologies (large lesions and deformations) to nevertheless be well registered with normal-control atlases. However, our current iterative registration framework needs to be better evaluated on TBI data sets with ground truth tissue labels. A near-term extension of this work is to form an unbiased atlas without a reference healthy atlas image, which is useful when only data containing pathologies is available. Future work will also focus on the development of a lesion segmentation pipeline using the sparse image as a spatial and intensity prior and a non-greedy implementation.

Acknowledgements. This work was supported, in-part, by the NIBIB (R41EB015775), the NINDS (R41NS081792) and the NSF (EECS-1148870).

References

1. Chitphakdithai, N., Duncan, J.: Non-rigid registration with missing correspondences in preoperative and postresection brain images. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part I. LNCS, vol. 6361, pp. 367–374. Springer, Heidelberg (2010)
2. Irimia, A., Wang, B., Aylward, S., Prastawa, M., Pace, D., Gerig, G., Hovda, D., Kikinis, R., Vespa, P., Van Horn, J.: Neuroimaging of structural pathology and connectomics in traumatic brain injury: Toward personalized outcome prediction. NeuroImage: Clinical 1, 1–17 (2012)
3. Joshi, S., Davis, B., Jomier, M., Gerig, G.: Unbiased diffeomorphic atlas construction for computational anatomy. NeuroImage 23, 151–160 (2004)
4. Lin, Z., Chen, M., Ma, Y.: The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv preprint arXiv:1009.5055 (2010)
5. Niethammer, M., Hart, G., Pace, D., Vespa, P., Irimia, A., Van Horn, J., Aylward, S.: Geometric metamorphosis. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part II. LNCS, vol. 6892, pp. 639–646. Springer, Heidelberg (2011)
6. Peng, Y., Ganesh, A., Wright, J., Xu, W., Ma, Y.: RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. TPAMI 34(11), 2233–2246 (2012)
7. Prastawa, M., Bullitt, E., Gerig, G.: Simulation of brain tumors in MR images for evaluation of segmentation efficacy. Med. Image Anal. 13(2), 297–311 (2009)
8. Rohlfing, T., Zahr, N.M., Sullivan, E.V., Pfefferbaum, A.: The SRI24 multichannel atlas of normal adult human brain structure. Hum. Brain Mapp. 31(5), 798–819 (2010)
9. Wang, B., Prastawa, M., Awate, S., Irimia, A., Chambers, M., Vespa, P., Van Horn, J., Gerig, G.: Segmentation of serial MRI of TBI patients using personalized atlas construction and topological change estimation. In: ISBI (2012)

Optimized PatchMatch for Near Real Time and Accurate Label Fusion

Vinh-Thong Ta[1,2,3], Rémi Giraud[1,2,3], D. Louis Collins[4], and Pierrick Coupé[1,2]

[1] Univ. Bordeaux, LaBRI, UMR 5800, PICTURA, F-33400 Talence, France
[2] CNRS, LaBRI, UMR 5800, PICTURA, F-33400 Talence, France
[3] IPB, LaBRI, UMR 5800, PICTURA, F-33600 Pessac, France
[4] McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Canada

Abstract. Automatic segmentation methods are important tools for quantitative analysis of magnetic resonance images. Recently, patch-based label fusion approaches demonstrated state-of-the-art segmentation accuracy. In this paper, we introduce a new patch-based method using the PatchMatch algorithm to perform segmentation of anatomical structures. Based on an Optimized PAtchMatch Label fusion (OPAL) strategy, the proposed method provides competitive segmentation accuracy in near real time. During our validation on hippocampus segmentation of 80 healthy subjects, OPAL was compared to several state-of-the-art methods. Results show that OPAL obtained the highest median Dice coefficient (89.3%) in less than 1 sec per subject. These results highlight the excellent performance of OPAL in terms of computation time and segmentation accuracy compared to recently published methods.

Keywords: PatchMatch, Patch-based Segmentation, Hippocampus.

1 Introduction

Automatic segmentation methods are efficient tools to produce accurate and reliable measurement dedicated to quantitative analysis of Magnetic Resonance Images (MRI). Over the past years, several paradigms were proposed to achieve the challenging task of brain labeling. First, atlas-based methods involving nonlinear registration of a labeled atlas to the subject to be segmented were proposed [1]. Then, multi-templates warping techniques based on training library of manually labeled templates were introduced. Such methods fuse several similar training templates to achieve better segmentation [2–4]. Multi-template matching approaches demonstrated competitive segmentation accuracies at the expense of an important computational load resulting from multiple nonlinear registrations (*i.e.*, up to several hours). Recently, a nonlocal patch-based label fusion (PBL) strategy [5] has been proposed. Requiring only linear registration, PBL involves patch comparison where the weight assigned to each label depends on the similarity between the current patch and the training patch. The search of similar training patches is based on nonlocal strategy to better handle the inter-subject

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 105–112, 2014.

variability and to capture registration inaccuracies. In a limited computational time (*i.e.*, several minutes), this method achieves state-of-the-art segmentation accuracy. Consequently, since its introduction, PBL is intensively studied and many improvements have been proposed [6–9].

Despite these improvements, PBL still suffers from several limitations. First, the search for similar patches is computationally expensive. Although template preselection [5], patch preselection [5, 6] or multiscale strategies [7] have been proposed, an important amount of computation remains dedicated to find similar patches in the training library. Second, the use of preselection strategy can prevent finding the most similar patches. In fact, similar patches can be found in dissimilar training templates. By removing *a priori* relevant parts of the training library, these preselection approaches can lead to sub-optimal results. Third, in PBL a weight is assigned to a large number of training patches including dissimilar patches. Therefore, resources are uselessly dedicated to estimate negligible weights. Even worse, these dissimilar patches can decrease the segmentation accuracy [8]. Sparsity-based methods can limit this aspect at the expense of an important computational burden [8, 9]. These limitations may result in sub-optimal segmentations and make the current implementations computationally expensive.

In this paper, we introduce a new PBL method based on the PatchMatch (PM) algorithm [10] to address these limitations. Originally, the PM algorithm was introduced to efficiently find an approximate nearest neighbor (ANN) for all patch correspondences between two 2D images. This method is based on a cooperative and randomized strategy resulting in very low computational burden that enables real time image processing. Recently, PM has been used for super-resolution of cardiac MRI [11]. Here, we propose a new Optimized PAtchMatch Label fusion (OPAL) method for anatomical structures segmentation by extending the PM approach. Compared to previous PBL methods, OPAL produces segmentations in near real time thanks to the use of the PM scheme. Moreover, OPAL does not require any pre-selection since the search of the most similar patch is achieved over the entire training library leading to higher segmentation accuracy. Finally, by using a very low number of highly similar patches, OPAL limits the introduction of dissimilar patches during label fusion.

The main contributions of this work are: (1) Adaptation of the PM algorithm to label fusion for anatomical structure segmentation in 3D MRI. (2) Acceleration techniques including constrained initialization, parallel processing and optimized distance computation. (3) Validation of OPAL on hippocampus segmentation. (4) Comparison with several state-of-the-art results in terms of computational time and segmentation accuracy.

2 Methods and Materials

2.1 The PatchMatch Algorithm

The original PM algorithm [10] is a fast and efficient approach that computes patch correspondences between two 2D images (denoted A and B). The key

point of this method is that good matches can be propagated to the adjacent patches within an image. This method is based on three steps: initialization, propagation and random search steps. The initialization consists in randomly associating a neighbor for each patch in A with a patch in B to obtain an initial ANN field. The propagation step tries to improve the patch correspondences using the observation that when a patch located at $p = (x, y) \in A$ matches well with a patch located at $q = (x', y') \in B$ then the adjacent patches of $p \in A$ should match well with the adjacent patches of $q \in B$. The random search step consists in a random sampling around the current ANN to escape from local minima. These two later steps are performed iteratively in order to improve the patch correspondences.

2.2 Optimized PatchMatch Algorithm

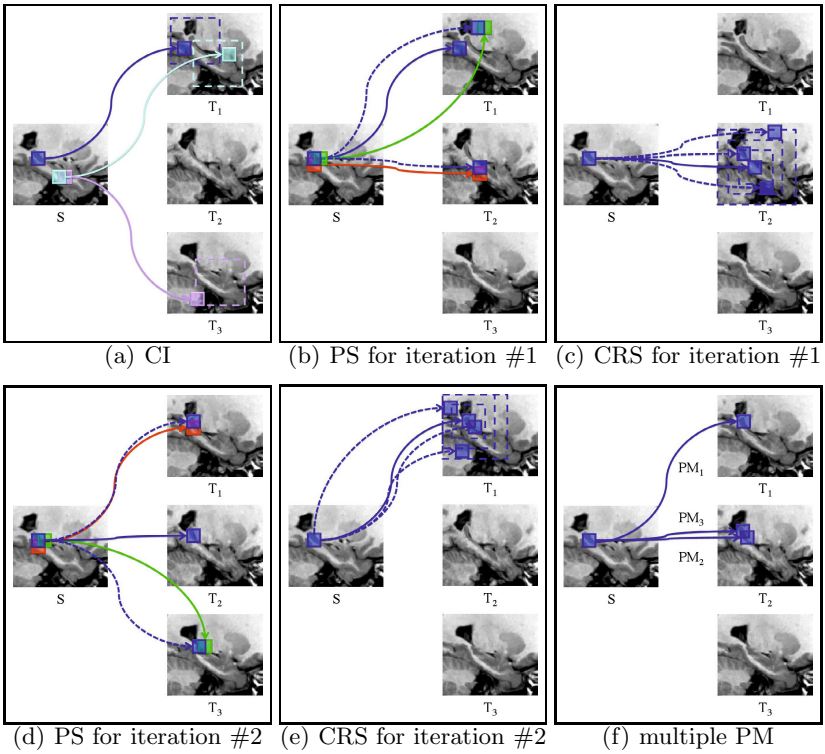
In contrast to [10] where two 2D images are considered, OPAL finds the patch correspondences between a 3D image S and a library of 3D templates $L = \{T_1, \ldots, T_n\}$ where n is the number of training templates. One advantage of the PM approach is that the complexity of this algorithm only depends on the size of image A and not on the size of the compared image B (*i.e.*, L in the OPAL case). This important fact allows OPAL to consider the entire image library L without any template preselection step at constant complexity. Moreover, for each patch in S, OPAL computes not only one match as done in [10] but the best k-ANN matches in L.

OPAL is explained in detail below while Figure 1 proposes its schematic overview. For the sake of clarity, only three templates are used in this figure and 3D MRI volumes are displayed in 2D instead of 3D.

Constrained Initialization. In [10], the initialization consists in affecting for each patch located at $(x, y) \in A$ a random correspondence located at $(x', y') \in B$. In the 3D case, the natural extension of this step would be to assign for each patch located at $(x, y, z) \in S$ a random patch correspondence located at $\{(x', y', z'), t\}$ where $t \in \{1, \ldots, n\}$ is the index of the template T_t within the library L. However, we can take advantage that all MRI volumes in L are linearly registered. Consequently, we propose to constrain the random initial position (x', y', z') to be within a fixed search window centered around the current position (x, y, z). Then, for each patch in S the index template t is assigned using *i.i.d.* random variable within $\{1, \ldots, n\}$. Figure 1(a) shows an illustration of this step, where for each patch in S (only three are displayed) the fixed search window for the random initialization is depicted in dotted lines in different training templates.

As in the PatchMatch algorithm, after this constrained initialization, propagation and random search steps are performed iteratively in order to improve the patch correspondence. Figure 1 also illustrates this iterative process.

Propagation Step with Fast Distance Computation. The OPAL propagation step extends the one proposed by [10] for the 3D case. For each patch located at $(x, y, z) \in S$, we try to improve its ANN by testing if the shifted ANN of its 6 adjacent patches located at $(x \pm 1, y, z)$, $(x, y \pm 1, z)$ and $(x, y, z \pm 1)$ provides a better match. Figures 1(b) and 1(d) illustrate this step, where the blue



(a) CI (b) PS for iteration #1 (c) CRS for iteration #1

(d) PS for iteration #2 (e) CRS for iteration #2 (f) multiple PM

Fig. 1. OPAL main steps. (a) Constrained initialization (CI), (b) and (d) propagation step (PS) for iteration #1 and #2, respectively (c) and (e) constrained random search (CRS) for iteration #1 and #2, respectively and (f) multiple PM. See text for more details.

dotted lines correspond to the test shifted adjacent neighbors in L in order to improve the current blue patch correspondence. In this example, the best match for the blue patch moves from template T_1 to T_2 with iteration #1 and from T_2 to T_1 with iteration #2. The propagation step is a core stage since it allows a patch correspondence to move over all the templates in L. Indeed, since the ANN of the adjacent voxels are not necessarily in the same template, the ANN of the current voxel can move from one template to another one.

Moreover, we propose an acceleration technique based on the observation that the ANN of the adjacent patches are known. Indeed, instead of computing the entire distance (the sum of the squared difference: SSD) between these patches, we take benefit from the patch overlapping by using a sliced SSD where only the non overlapping coordinates are considered. Finally, during the SSD computation, we test if the current sum is superior to the previous minimal SSD. By this way, the SSD estimation can be stopped avoiding extra computation.

Constrained Random Search. In contrast to [10], OPAL deals with a library of images. Therefore, we modify the random search step in order to take into

account this aspect. Indeed, if we use the original PM algorithm, the random search step should be performed on all the L dimensions, *i.e.*, x, y, z and t. However, to ensure spatial consistency, OPAL performs the random search only in the current template that provides the current best patch correspondence (*i.e.*, t is fixed, and we random on $(x', y', z') \in T_t$) within a decaying search window as in [11]. Figures 1(c) and 1(e) present examples of such fixed template random search where the decaying search windows are represented in dotted blue lines.

Multiple PM and Parallel Computation. Finally, while in [10] only the best match is estimated, OPAL computes k-ANN matches in L to perform label fusion. In the literature, an extension of the original PM algorithm to k-ANN case was proposed [12]. The suggested strategy is to build a stack of the best visited matches to obtain the k-ANN. However, to parallelize such an approach, the current image S must be split into several parts with problems of patch boundaries overlapping between threads. Therefore, in OPAL, we decided to based the k-ANN search on independent k-PM enabling a more efficient and simple multi-threading. Figure 1(f) illustrates the result of the multiple PM step where here $k = 3$ and, each $\text{PM}_{i=1,2,3}$ denotes an individual PM.

Patch-Based Label Fusion. At the end the process, the k-ANN are estimated for all the patches in S. Thus, the location and the SSD between the patches of S and their k-ANN in L are known. Therefore, to obtain the final segmentation, we used the PBL method presented in [5]. However, in OPAL only the k most similar patches are used (limiting segmentation error) and the entire library is considered (increasing segmentation accuracy). When the same ANN is selected several times it will be considered several times during label fusion. Finally, to further improve segmentation quality, label fusion is performed over the whole patch as done in [6, 9] and not only using the central voxel.

2.3 Validation

Dataset. The proposed method was evaluated on the International Consortium for Brain Mapping (ICBM) dataset. Part of this dataset consists of 80 MR images of young and healthy individuals with manual segmentations following the Pruessner's protocol [13]. The MRI scans were acquired with a 1.5T Philips GyroScan imaging system (1 mm thick slices, TR = 17 ms, TE = 10 ms, flip angle = 30 °, 256 mm field of view). The estimated intra-class reliability coefficient was of 0.90 for inter- (4 raters) and 0.92 for intra-rater (5 repeats) reliability.

Preprocessing. All the images were preprocessed through the following pipeline: estimation of the standard deviation of noise [14]; denoising using the optimized nonlocal means filter [15]; correction of inhomogeneities using N3 [16]; registration to stereotaxic space based on a linear transform to the ICBM152 template ($1\times1\times1$ mm^3 voxel size) [17]; linear intensity normalization of each subject on template intensity; brain extraction using BEaST [7]; image cropping around the structures of interest; and cross-normalization of the MRI intensity between the subjects within the estimated brain mask with [18].

Quality Metric and Compared Methods. To validate the proposed method a leave-one-out cross validation procedure was used. During our validation,

we investigated the impact of the patch size and of the number of neighbors (*i.e.*, number of PM). Moreover, OPAL was compared with Atlas-Based Method (ABM) [1], Multi-Templates Matching (MTM) [4], Patch-Based Label fusion (PBL) [5], Sparse Representation Classification (SRC) [8], Discriminative Dictionary Learning for Segmentation (DDLS) and Fixed Discriminative Dictionary Learning for Segmentation (F-DDLS) [8] since all these methods were validated on the same dataset. The segmentation quality was estimated with the Dice coefficient by comparing the expert-based segmentations with the automatic segmentations. The median Dice coefficients and computational times presented in Table 1 are the published values. These values include segmentation of both, left and right hippocampus. OPAL was implemented in MATLAB using multi-threaded C-MEX code. Our experiments were carried out using a server of 16 cores at 2.6 GHz with 100 GB of RAM. The number of threads was equal to k and the number of inner iterations of OPAL was set to 5 as in [10].

3 Results

Influence of Parameters. Figure 2 shows the influence of the number of neighbors and of the patch size on the segmentation quality and on the computational time. Similarly to previous PBL methods [5, 8], we found that patches of size 5×5×5 and 7×7×7 voxels provides the best results with a slight advantage for patches of size 5×5×5 voxels (89.4% for $k = 20$). Moreover, we found that the median Dice coefficient reached a plateau around 10-ANN. Interestingly, this number is in line with the suggested number of templates in multi-templates matching methods [4]. As expected, bigger patches and larger number of ANN required higher computational time. Consequently, our experiments suggest that using patch size of 5×5×5 and $k = 10$ offers a good trade off between segmentation accuracy (89.3%) and computational time (0.89s).

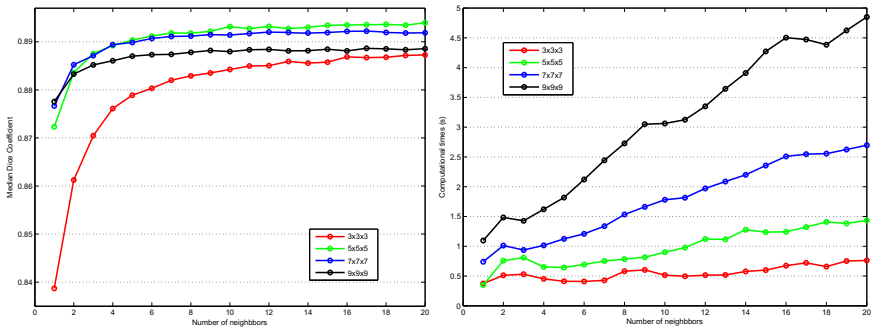


Fig. 2. Median Dice coefficient according to the patch size and the number of neighbors (at left) and the corresponding computational time (at right)

Comparison with State-of-the-Art Methods. The comparison of OPAL performance with 6 other methods is presented in Table 1. The presented values

Table 1. Methods comparison in terms of segmentation accuracy and computational time on the ICBM dataset

Method	Median Dice Coefficient	Computational Time by Subject
ABM [1]	86.4%	358s
PBL [5]	88.2%	662s
MTM [4]	88.6%	3974s
F-DDLS [8]	88.6%	193s
SRC [8]	88.7%	5587s
DDLS [8]	89.0%	943s
OPAL	**89.3**%	**0.89s**

are the results published by the authors for the segmentation of both hippocampi on the ICBM dataset. The provided computational times do not include template preselection while only OPAL does not require it. Therefore, the computational times are under-estimated except for OPAL. Moreover, for the F-DDLS an offline training step of 1781s is needed. However, OPAL obtained the highest median Dice coefficient in the fastest manner. These results highlights the excellent performance of OPAL in terms of both: segmentation accuracy and computational time. Compared to the original PBL [5], OPAL obtained better accuracy 700× faster. Moreover, OPAL obtained the highest Dice coefficient for a computational time 200× faster than the fastest published method on the used dataset (F-DDLS [8]). Finally, compared to the most accurate method (DDLS [8]), OPAL obtained higher Dice coefficient for a computational time 1000× faster.

4 Conclusion

In this paper, we propose a novel patch-based segmentation method based on an optimized PatchMatch label fusion. The Opal method enables high quality segmentation in near real time. Experiments show that the proposed method obtained competitive results compared to the state-of-the-art approaches. Indeed, the OPAL obtained the highest median Dice coefficient in a much faster manner. In addition, the near real time capabilities of OPAL pave the way for new applications for label fusion segmentation. For instance, OPAL can be used as an efficient automatic or interactive segmentation tool in medical visualization software. Finally, as future work, OPAL will be validated on multi-sites datasets containing pathological cases and extended to multi-label segmentation.

Acknowledgments. This study has been carried out with financial support from the French State, managed by the French National Research Agency (ANR) in the frame of the Investments for the future Programme IdEx Bordeaux (ANR-10-IDEX-03-02), Cluster of excellence CPU and TRAIL (HR-DTI ANR-10-LABX-57). We also acknowledge funding from the Fonds de Recherche Québec- Santé (FRQS-Pfizer).

References

1. Collins, D.L., et al.: Automatic 3-D model-based neuroanatomical segmentation. Human Brain Mapping 3(3), 190–208 (1995)
2. Heckemann, R.A., et al.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 33(1), 115–126 (2006)
3. Lötjönen, J.M., et al.: Fast and robust multi-atlas segmentation of brain magnetic resonance images. Neuroimage 49(3), 2352–2365 (2010)
4. Collins, D.L., Pruessner, J.C.: Towards accurate, automatic segmentation of the hippocampus and amygdala from mri by augmenting animal with a template library and label fusion. Neuroimage 52(4), 1355–1366 (2010)
5. Coupé, P., et al.: Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage 54(2), 940–954 (2011)
6. Rousseau, F., Habas, P.A., Studholme, C.: A supervised patch-based approach for human brain labeling. IEEE Transactions on Medical Imaging 30(10), 1852–1862 (2011)
7. Eskildsen, S.F., et al.: BEaST: Brain extraction based on nonlocal segmentation technique. NeuroImage 59(3), 2362–2373 (2012)
8. Tong, T., et al.: Segmentation of MR images via discriminative dictionary learning and sparse coding: Application to hippocampus labeling. NeuroImage 76, 11–23 (2013)
9. Wu, G., et al.: A generative probability model of joint label fusion for multi-atlas based brain segmentation. Medical Image Analysis (2013)
10. Barnes, C., et al.: Patchmatch: A randomized correspondence algorithm for structural image editing. In: ACM SIGGRAPH 2009 Papers, pp. 24:1–24:11 (2009)
11. Shi, W., et al.: Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part III. LNCS, vol. 8151, pp. 9–16. Springer, Heidelberg (2013)
12. Barnes, C., Shechtman, E., Goldman, D.B., Finkelstein, A.: The generalized patchmatch correspondence algorithm. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 29–43. Springer, Heidelberg (2010)
13. Pruessner, J., et al.: Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. Cerebral Cortex 10(4), 433–442 (2000)
14. Coupé, P., et al.: Robust Rician noise estimation for MR images. Medical Image Analysis 14(4), 483–493 (2010)
15. Coupé, P., et al.: An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. IEEE Transactions on Medical Imaging 27(4), 425–441 (2008)
16. Sled, J.G., et al.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Transactions on Medical Imaging 17(1), 87–97 (1998)
17. Collins, D.L., et al.: Automatic 3D intersubject registration of MR volumetric data in standardized talairach space. Journal of Computer Assisted Tomography 18(2), 192–205 (1994)
18. Manjón, J.V., et al.: Robust MRI brain tissue parameter estimation by multistage outlier rejection. Magnetic Resonance in Medicine 59(4), 866–873 (2008)

Functionally Driven Brain Networks Using Multi-layer Graph Clustering

Yasser Ghanbari[1], Luke Bloy[2], Varsha Shankar[1], J. Christopher Edgar[2], Timothy P.L. Roberts[2], Robert T. Schultz[2], and Ragini Verma[1,⋆]

[1] Center for Biomedical Image Computing and Analytics, University of Pennsylvania
{Yasser.Ghanbari,Varsha.Shankar,Ragini.Verma}@uphs.upenn.edu
[2] Center for Autism Research, Children's Hospital of Philadelphia, Philadelphia, PA
{bloyl,edgarj,robertstim,schultzrt}@chop.edu

Abstract. Connectivity analysis of resting state brain has provided a novel means of investigating brain networks in the study of neurodevelopmental disorders. The study of functional networks, often represented by high dimensional graphs, predicates on the ability of methods in succinctly extracting meaningful representative connectivity information at the subject and population level. This need motivates the development of techniques that can extract underlying network modules that characterize the connectivity in a population, while capturing variations of these modules at the individual level. In this paper, we propose a multi-layer graph clustering technique that fuses the information from a collection of connectivity networks of a population to extract the underlying common network modules that serve as network hubs for the population. These hubs form a functional network atlas. In addition, our technique provides subject-specific factors designed to characterize and quantify the degree of intra- and inter- connectivity between hubs, thereby providing a representation that is amenable to group level statistical analyses. We demonstrate the utility of the technique by creating a population network atlas of connectivity by examining MEG based functional connectivity in typically developing children, and using this to describe the individualized variation in those diagnosed with autism spectrum disorder.

1 Introduction

Computational techniques applied to neuroimaging data have helped characterize brain connectivity anomalies in autism spectrum disorder (ASD) and schizophrenia. While structural connectivity is based on tractography using diffusion MRI [1], functional connectivity is investigated by using coherence measures between regions [2] using fMRI or magnetoencephalography (MEG) [3].

The study of brain connectivity networks has recently gained considerable attention. The high dimensionality of these networks as well as their variation at the subject level within the population calls for methods that can elucidate the

⋆ Authors acknowledge support from grants NIH-MH098010 (PI: R. Verma), Pennsylvania Department of Health, SAP#4100042728, SAP#4100047863 (PI: R. Schultz).

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 113–120, 2014.

underlying network structure while reducing dimensionality. In this paper, we present a novel approach that extracts the underlying network modules that describe the hubs of brain connectivity. Such modules are characterized by highly inter-connected regions within the module, in comparison to their connectivity to regions outside the module. This collection of network modules can serve as an atlas of network variation in a population. In addition to extracting these neurophysiological network hubs, our method provides representations of intra- and inter-connectivity strength of these network hubs for each subject, facilitating group-based statistical analysis.

The approach we take to extract these network hubs is based on multi-layer graph clustering. The advent of graph-based clustering techniques has led to recent growing interest in methods for clustering of multi-layer graphs in the area of mobile phone networks and document clustering [4–6]. However, such methods are mainly concerned with the approximation of graph Laplacian to feed the spectral clustering algorithm, and lack interpretability. In our approach, we present a framework for multi-layer graph clustering for analysis of connectivity in terms of splitting the brain network into hubs characteristic of a population and their low-dimensional interaction weights amenable to group-wise statistics. The connectivity network of each typically developing control (TDC) is represented as a graph, and all TDC graphs are stacked to form a multi-layer graph, each layer representing an individualized variation of the underlying network connectivity. A matrix factorization model is employed to decompose the set of healthy connectivity graphs into clusters of network modules (hubs) shared among all graph layers. These network hubs are learned by using a gradient descent approach minimizing the reconstruction error of decomposition in the healthy population network set. The network hubs obtained are then used adaptively to optimize hubs intra- and inter-connectivity weights for each subject.

While our method is generalizable to non-negative connectivity matrices obtained from DTI or fMRI, we demonstrate its applicability to resting-state MEG connectivity networks in alpha frequency-band for a population of ASD subjects.

2 Methods

Our framework is based on the premise that there are a few underlying subnetworks that describe a population with variation demonstrated between each subject. The method we present here determines the network hubs that characterize the commonality across all subjects within a population (e.g. default mode network), with the interaction within and between these hubs that captures the individualized variation in each subject. Therefore, we capture not only the dominant network hubs that describe a population, but also the subtle connectivity between these hubs that describes the variation in each subject either due to inherent heterogeneity or induced by pathology. This collection of network hubs will be referred to as the network atlas.

Given a population, we create this network atlas using the connectivity matrices from all the subjects. The connectivity is quantified by a non-negative similarity measure between n regions leading to a non-negative connectivity matrix of

subject m, i.e. $\boldsymbol{S}^{(m)} \in \mathbb{R}^{n\times n}$, represented by a graph with n vertices. We then use a matrix factorization model $\boldsymbol{S}^{(m)} \approx \boldsymbol{U}\boldsymbol{\Lambda}^{(m)}\boldsymbol{U}^T$ where $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k] \in \mathbb{R}^{n\times k}$ is the common factor of the population characterizing the shared underlying connectivity modules of the population. $\boldsymbol{\Lambda}^{(m)} = \left[\lambda_{ij}^{(m)}\right] \in \mathbb{R}^{k\times k}$ is also the symmetric subject-level factor capturing the weights of each subject's network modules. $k \ll n$ is the number of network hubs (modules) to be identified. Due to the symmetry of $\boldsymbol{\Lambda}^{(m)}$, this decomposition model can be re-written as

$$\boldsymbol{S}^{(m)} \approx \boldsymbol{U}\boldsymbol{\Lambda}^{(m)}\boldsymbol{U}^T = \sum_{i=1}^{k} \lambda_{ii}^{(m)} \boldsymbol{u}_i \boldsymbol{u}_i^T + \sum_{i=1}^{k}\sum_{\substack{j=1 \\ j>i}}^{k} \lambda_{ij}^{(m)} \left(\boldsymbol{u}_i \boldsymbol{u}_j^T + \boldsymbol{u}_j \boldsymbol{u}_i^T\right). \tag{1}$$

In this model, each network hub is identified by the first term in (1), i.e. $\boldsymbol{u}_i\boldsymbol{u}_i^T$, whose subject-level intra-connectivity strength is represented by coefficients $\lambda_{ii}^{(m)}$. On the other hand, the subject-level inter-connectivity strength between hubs i and j is represented by $\lambda_{ij}^{(m)}$ where the inter-connectivity pattern is identified by $\boldsymbol{u}_i\boldsymbol{u}_j^T + \boldsymbol{u}_j\boldsymbol{u}_i^T$, i.e. the second term of (1). Elements of $\boldsymbol{U}$ are constrained to remain non-negative $U_{ij} \geq 0$, thus retain the interpretation of its components (i.e. $\boldsymbol{u}_i\boldsymbol{u}_j^T$) as a connectivity matrix (i.e. hubs and their inter-connectivity modules). $\boldsymbol{\Lambda}^{(m)}$ is constrained to be non-negative $\lambda_{ij}^{(m)} \geq 0$ and symmetric $\lambda_{ij}^{(m)} = \lambda_{ji}^{(m)}$ due to the symmetry of connectivity matrices $\boldsymbol{S}^{(m)}$, but no constraints are imposed on it to be diagonal, because this lets our model capture the inter- connectivity weights on off-diagonal elements, and not overlook the interactions between network hubs in the study.

2.1 Objective Function

Since we would like to obtain the underlying network modules shared among all subjects in the population, we stack the connectivity graph of all subjects to form a multi-layer graph $\left\{\boldsymbol{S}^{(m)}\right\}$. The network hubs shared by the population is then obtained by minimizing the reconstruction error of the decomposition across layers. This can be obtained by minimizing the following objective function with appropriate constraints on $\boldsymbol{U}$ and $\boldsymbol{\Lambda}^{(m)}$ as explained in equation (1),

$$\begin{aligned} &J\left(\boldsymbol{U}, \boldsymbol{\Lambda}^{(m)}\right) = \sum_{m=1}^{M} \|\boldsymbol{S}^{(m)} - \boldsymbol{U}\boldsymbol{\Lambda}^{(m)}\boldsymbol{U}^T\|_F^2 + \beta\left(\|\boldsymbol{U}\|_F^2 + \sum_{m=1}^{M}\|\boldsymbol{\Lambda}^{(m)}\|_F^2\right), \\ &\text{subject to } \; U_{ij} \geq 0 \;\; , \;\; \lambda_{ij}^{(m)} \geq 0 \;\text{ and }\; \boldsymbol{\Lambda}^{(m)} = \boldsymbol{\Lambda}^{(m)^T} \end{aligned} \tag{2}$$

where M is the number of subjects within the population, and $\|.\|_F$ denotes the Frobenius norm. The regularization term, as the sum of the squared norm of $\boldsymbol{U}$ and $\boldsymbol{\Lambda}^{(m)}$, is added to improve the numerical stability, and β is a tunable parameter balancing the two terms of reconstruction error norm and regularization.

2.2 Optimization Solution

Due to symmetry of $\boldsymbol{S}^{(m)}$ and $\boldsymbol{\Lambda}^{(m)}$, equation (2) can be rewritten as

$$\begin{aligned} J\left(\boldsymbol{U}, \boldsymbol{\Lambda}^{(m)}\right) = & \sum_{m=1}^{M} trace\left\{\left(\boldsymbol{S}^{(m)} - \boldsymbol{U}\boldsymbol{\Lambda}^{(m)}\boldsymbol{U}^T\right)^2\right\} \\ & + \beta\left(trace\left\{\boldsymbol{U}\boldsymbol{U}^T\right\} + \sum_{m=1}^{M} trace\left\{\boldsymbol{\Lambda}^{(m)^2}\right\}\right), \\ & \text{subject to } U_{ij} \geq 0 \ , \ \lambda_{ij}^{(m)} \geq 0 \text{ and } \boldsymbol{\Lambda}^{(m)} = \boldsymbol{\Lambda}^{(m)^T}. \end{aligned} \tag{3}$$

To minimize (3), we propose an iterative procedure in which $\boldsymbol{U}$ and $\boldsymbol{\Lambda}^{(m)}$ are alternately optimized by given multi-layer graph of the population $\left\{\boldsymbol{S}^{(m)}\right\}$. We use the gradient decent approach, i.e. alternately updating $U_{ij} = U_{ij} - \eta_{ij}\frac{\partial J}{\partial U_{ij}}$ and $\lambda_{ij}^{(m)} = \lambda_{ij}^{(m)} - \zeta_{ij}^{(m)}\frac{\partial J}{\partial \lambda_{ij}^{(m)}}$ with step-sizes $\eta_{ij} \geq 0$ and $\zeta_{ij}^{(m)} \geq 0$, where

$$\frac{\partial J}{\partial \boldsymbol{U}} = -4\sum_{m=1}^{M}\left[\left(\boldsymbol{S}^{(m)} - \boldsymbol{U}\boldsymbol{\Lambda}^{(m)}\boldsymbol{U}^T\right)\boldsymbol{U}\boldsymbol{\Lambda}^{(m)}\right] + 4\beta\boldsymbol{U}, \tag{4}$$

$$\frac{\partial J}{\partial \boldsymbol{\Lambda}^{(m)}} = -2\boldsymbol{U}^T\left(\boldsymbol{S}^{(m)} - \boldsymbol{U}\boldsymbol{\Lambda}^{(m)}\boldsymbol{U}^T\right)\boldsymbol{U} + 2\beta\boldsymbol{\Lambda}^{(m)}. \tag{5}$$

Due to non-negativity of $\boldsymbol{S}^{(m)}$, our non-negativity constraints will be guaranteed by positive initialization of $\boldsymbol{U}$ and (symmetric) $\boldsymbol{\Lambda}^{(m)}$, and applying the step sizes as $\eta_{ij} = \frac{\frac{1}{4}U_{ij}}{\left(\beta\boldsymbol{U}+\sum_{m=1}^{M}\boldsymbol{U}\boldsymbol{\Lambda}^{(m)}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Lambda}^{(m)}\right)_{ij}}$, $\zeta_{ij}^{(m)} = \frac{\frac{1}{2}\lambda_{ij}^{(m)}}{\left(\beta\boldsymbol{\Lambda}^{(m)}+\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Lambda}^{(m)}\boldsymbol{U}^T\boldsymbol{U}\right)_{ij}}$. This results in the following multiplicative updating solutions

$$U_{ij} = U_{ij}\frac{\left(\sum_{m=1}^{M}\boldsymbol{S}^{(m)}\boldsymbol{U}\boldsymbol{\Lambda}^{(m)}\right)_{ij}}{\left(\beta\boldsymbol{U} + \sum_{m=1}^{M}\boldsymbol{U}\boldsymbol{\Lambda}^{(m)}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Lambda}^{(m)}\right)_{ij}}, \tag{6}$$

$$\lambda_{ij}^{(m)} = \lambda_{ij}^{(m)}\frac{\left(\boldsymbol{U}^T\boldsymbol{S}^{(m)}\boldsymbol{U}\right)_{ij}}{\left(\beta\boldsymbol{\Lambda}^{(m)} + \boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Lambda}^{(m)}\boldsymbol{U}^T\boldsymbol{U}\right)_{ij}}, \ \text{ for } 1 \leq m \leq M. \tag{7}$$

Starting with initial random positive elements on $\boldsymbol{U}$ and (symmetric) $\boldsymbol{\Lambda}^{(m)}$, the iterative procedures (6) and (7) are performed alternately until convergence. Such initialization will guarantee non-negativity on $\boldsymbol{U}$ and $\boldsymbol{\Lambda}^{(m)}$ as well as the symmetry of $\boldsymbol{\Lambda}^{(m)}$, which can be verified from equations (6) and (7).

2.3 Optimizing Subject-Level Factors

The above procedure is performed on the multi-layer graph of a population to create an atlas of network hubs. The common factor $\boldsymbol{U}$ obtained from them is employed to optimize equation (2) to compute subject-level factors $\boldsymbol{\Lambda}^{(m)}$ from

their connectivity matrices. Since the common factor $\boldsymbol{U}$ is given, (2) can be minimized by only performing the iterative procedure of (7) with random symmetric non-negative initialization of $\boldsymbol{\Lambda}^{(m)}$ for each subject, yielding the subject-specific intra- and inter-connectivity weight of network hubs. It is worth noting that, given $\boldsymbol{U}$, equation (7) is iterated on both control and patient data so that both populations undergo the same procedure to be comparable in statistics.

2.4 Statistical Analysis and Interpretation

As explained at the beginning of section 2, elements of the subject-level factors $\boldsymbol{\Lambda}^{(m)}$ represent weights of network hubs in that subject. The intra-connectivity of network hubs is represented by the diagonal elements of $\boldsymbol{\Lambda}^{(m)}$, i.e. $\lambda_{ii}^{(m)}$, and their inter-connectivity is represented by the upper triangular elements (due to symmetry of it). Hence, a significant group difference at a diagonal element $\lambda_{ii}^{(m)}$ interprets as an alteration in the communication within the i^{th} network hub, characterized by $\boldsymbol{u}_i\boldsymbol{u}_i^T$, and a group difference at an off-diagonal element $\lambda_{ij}^{(m)}$ indicates changes in the interaction between the network hubs i and j, i.e. the inter-connectivity pattern $\boldsymbol{u}_i\boldsymbol{u}_j^T + \boldsymbol{u}_j\boldsymbol{u}_i^T$ has been affected by disease.

3 Results

The proposed method provides a means to extract common population level information while also capturing individual subject variation. We demonstrate the applicability of this methodology to the study of resting-state (RS) MEG functional connectivity in a population of ASD, in comparison to a network hub atlas that has been learned on a TDC population. The connectivity of the entire population is then represented in this more concise module representation.

In order to perform a preliminary test, we applied our method to a set of synthetic noisy networks with known common modules (hubs), which were successfully restored by the procedure explained in section 2.

Dataset Demographics. The dataset consisted of 77 male children, 37 ASD and 40 TDCs, aged 6-14 years (mean=10.2 years, SD=1.8 in ASD, and mean=10.3 years, SD=1.7 in TDC, no significant difference in age $p > 0.6$). RS MEG was acquired in a magnetically shielded room using a 306-channel Elekta scanner. Two minutes of recorded data were obtained after artifact removal, which were low-pass filtered before downsampling to 500 Hz to avoid aliasing.

Source Localization and Connectivity Measures. RS eyes-closed data were band-pass filtered to the alpha band activity (8–12 Hz). MEG data were divided into 2.5-second epochs with 50% overlap and transformed into the frequency domain. A 5mm isotropic source grid was obtained by sampling cortical gray-matter areas from the T1-weighted MRI of each subject. The sensor-space frequency-domain data were used by VESTAL [7] to obtain source amplitude at each source location. From this spatial distribution of source amplitudes, an inverse operator was determined [7] and applied to the MEG data yielding source time-courses at each location. 301 structurally meaningful regions of interest

(ROIs) were determined using Freesurfer tools to subdivide the cortical surface of a template subject and to map these ROIs into each of the 77 subjects. Of the 301 ROIs, we identified 202 ROIs that have at least one source assigned to them. ROI dominant time-courses were then determined by using singular value decomposition. Connectivity matrices were computed for the 202 regions yielding 77 matrices of size 202×202. Synchronization likelihood (SL), a non-negative measure of synchronous activity between 0 (no connection) and 1 (completely synchronous), was used to quantify the connectivity between two regions [3].

Connectivity Network Analysis. The SL connectivity matrices of 40 TDC subjects were used to compute the network hubs as well as their 40 subject-level weights. We set $\beta = 0.1$ and used $k = 10$ to obtain 10 network hubs. The iterative procedure of (6) and (7) was performed that converged to the network atlas $\boldsymbol{U}$ of size 202×10. The resulting ten network hubs (i.e. the first term in equation (1), $\boldsymbol{u}_i \boldsymbol{u}_i^T$ for $1 \leq i \leq 10$) are shown together on a brain map in Fig. 1, thresholded for binary visualization of their dominant edges.

Given the network hubs computed for the TDC population, the subject-level weights $\boldsymbol{\Lambda}^{(m)}$, each of size 10×10, for the 40 TDC as well as 37 ASD connectivity matrices were obtained. These subject-level weights of the two groups were then used to perform statistical group comparison for each of 10 diagonal and 45 upper-triangular elements of $\boldsymbol{\Lambda}$ across subjects, out of which five of upper-triangular elements indicated significant group differences ($p < 0.05$). In Fig. 2, we show the five inter-connectivity patterns that correspond to the significant weights, by displaying their inter-connectivity maps generated by the second term of equation (1), i.e. $\boldsymbol{u}_i \boldsymbol{u}_j^T + \boldsymbol{u}_j \boldsymbol{u}_i^T$, as explained in section 2.4. The inter-connectivity patterns which have larger weights in TDC are shown in blue (Fig. 2 (a) to (c)), while those larger in ASD are shown in orange (Fig. 2 (d) and (e)).

It is observed from Fig. 1 that our method has extracted modular brain functional hubs that are spatially close but sparsely distributed on the cortex. The arrangement of these network hubs may also define the default functional network in MEG brain connectivity. As mentioned above, the statistical group comparison of ten diagonal elements of subject-level weights did not show significant difference (with $p < 0.05$) indicating that the contribution of network hubs does not substantially differ between ASD and TDCs. It is however worth noting that Hub#5 demonstrated the most different intra-connectivity among the ten hubs shown in Fig. 1. This network showed a group difference of $p < 0.1$ (with higher weights in ASD) indicating hyperconnectivity in short-range communications in the very frontal area of ASD brain, an observation consistent with the literature of ASD frontal lobe overconnectivity [8]. It is interesting to observe that such enhanced connectivity within this hub in ASD, coincides with underconnectivity in the interaction of this hub with bilateral hubs Hub#1 and Hub#6 (Fig. 2(a) and (c)) as well as the occipital hub Hub#4 (Fig. 2(b)). This may be an indication that the local overconnectivity in the front of ASD brain has led to long-distance underconnectivity, or vice versa.

Fig. 2 (a)–(c) shows that ASD brain has deficient long-range connectivity primarily in fronto-occipital communications (connectivity pattern between Hub#4

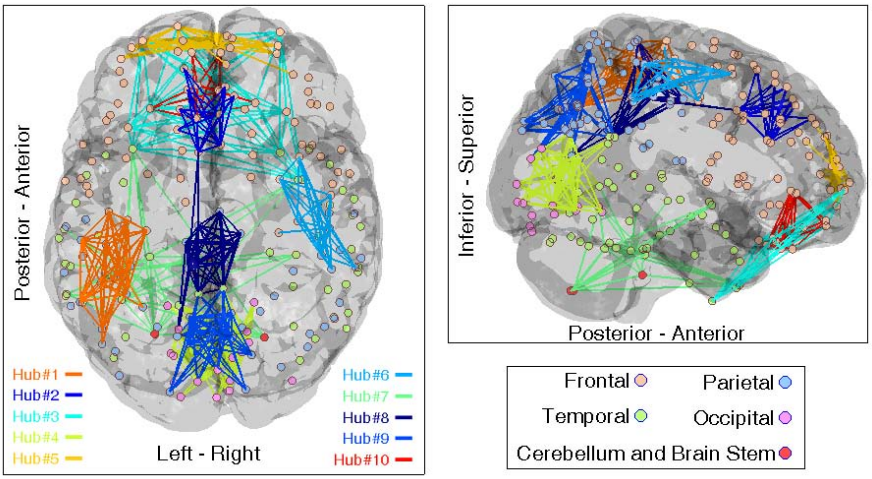


Fig. 1. The k=10 functional network hubs of brain MEG alpha activity obtained from the 40 TDC connectivity matrices, displayed together with no specific order

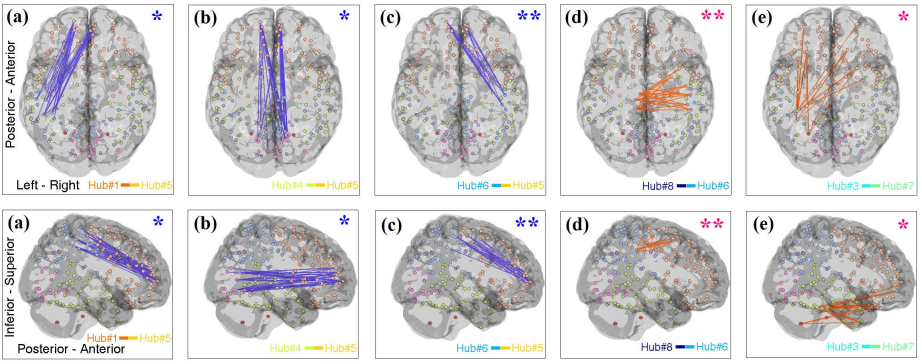


Fig. 2. The five inter-connectivity (between-hub) patterns with significant group difference between ASD and TDC (axial and sagittal view). Blue and orange inter-connectivity networks have higher weights in TDC and ASD, respectively. $p < 0.05$ and $p < 0.01$ are labelled with one and two stars, respectively.

and Hub#5 shown in Fig. 2 (b)) as well as in fronto-parietal connections (between Hub#1 and Hub#5 as well as Hub#6 and Hub#5 shown respectively in Fig. 2 (a) and (c)), consistent with findings of fronto-posterior under-connectivity in autism [9]. In addition, Fig. 2(d) shows that ASD has enhanced short-range connectivity (between Hub#8 and Hub#6). Fig. 2(e) also shows increased inter-connectivity between the frontal and temporal/subcortical regions in ASD (between Hub#3 and Hub#7).

The network hubs obtained here are clinically interpretable as they implicate the regions associated with the known default mode network (DMN) which is

the network most commonly elucidated in functional resting state literature. We have also observed that the average weights of the intra-connectivity networks (i.e. hubs) are an order of magnitude larger than the inter-connectivity weights. Therefore, the inter-connectivity network modules do not substantially contribute to reconstructing subjects' functional networks, but important in characterizing the effect of disease. Finally, we tested the method on subsets of healthy subjects and have obtained similar network hubs indicating the repeatability of the results.

4 Conclusion

We have presented a new analysis technique of connectivity matrices using a low-rank matrix factorization model that extracts a set of population specific network hubs shared by all the matrices in a multi-layer graph framework. Application to a dataset of TDC subjects provided a set of functional network hubs and their intra- and inter-connectivity weights. The network hubs obtained from TDC were used to compute the subject-level weights for ASD subjects. Group-wise analysis of intra- and inter-connectivity weights revealed significant long-range connectivity deficiencies as well as short-range overconnectivity in ASD. The proposed framework can be extended to any non-negative connectivity matrix, and the weights obtained in the process can be exploited for statistical analysis.

References

1. Basser, P.J., Pajevic, S., et al.: In vivo fiber tractography using dt-mri data. Magn. Reson. Med. 44(4), 625–632 (2000)
2. Vissers, M., et al.: Brain connectivity and high functioning autism: a promising path of research that needs refined models, methodological convergence, and stronger behavioral links. Neurosci. Biobehav. Rev. 36(1), 604–625 (2012)
3. Ghanbari, Y., Bloy, L., Edgar, J.C., et al.: Joint analysis of band-specific functional connectivity and signal complexity in autism. J. Autism. Dev. Disord. (2013)
4. Dong, X., Frossard, P., et al.: Clustering on multi-layer graphs via subspace analysis on grassmann manifolds. IEEE Trans. Sig. Proc. 62, 905–918 (2014)
5. Dong, X., Frossard, P., et al.: Clustering with multi-layer graphs: A spectral perspective. IEEE Trans. Sig. Proc. 60, 5820–5831 (2012)
6. Tang, W., Lu, Z., Dhillon, I.: Clustering with multiple graphs. In: IEEE Int. Conf. Data Mining, Miami, Fl, pp. 1016–1021 (2009)
7. Huang, M.X., Huang, C.W., Robb, A., et al.: Meg source imaging method using fast l1 minimum-norm and its applications to signals with brain noise and human resting-state source amplitude images. Neuroimage 84, 585–604 (2014)
8. Kana, R.K., Libero, L.E., et al.: Disrupted cortical connectivity theory as an explanatory model for autism spectrum disorders. Phys. Life Rev. 8, 410–437 (2011)
9. Just, M.A., Keller, T.A., et al.: Autism as a neural systems disorder: a theory of frontal-posterior underconnectivity. Neurosci. Biobehav. Rev. 36, 1292–1313 (2012)

Bayesian Principal Geodesic Analysis in Diffeomorphic Image Registration

Miaomiao Zhang and P. Thomas Fletcher

School of Computing, University of Utah, Salt Lake City, USA

Abstract. Computing a concise representation of the anatomical variability found in large sets of images is an important first step in many statistical shape analyses. In this paper, we present a generative Bayesian approach for automatic dimensionality reduction of shape variability represented through diffeomorphic mappings. To achieve this, we develop a latent variable model for principal geodesic analysis (PGA) that provides a probabilistic framework for factor analysis on diffeomorphisms. Our key contribution is a Bayesian inference procedure for model parameter estimation and simultaneous detection of the effective dimensionality of the latent space. We evaluate our proposed model for atlas and principal geodesic estimation on the OASIS brain database of magnetic resonance images. We show that the automatically selected latent dimensions from our model are able to reconstruct unseen brain images with lower error than equivalent linear principal components analysis (LPCA) models in the image space, and it also outperforms tangent space PCA (TPCA) models in the diffeomorphism setting.

1 Introduction

Diffeomorphic image registration plays an important role in understanding anatomical shape variability in medical image analysis. For example, analysis of diffeomorphic shape changes can be linked to disease processes and changes in cognitive and behavioral measures. In this setting, the high dimensionality of the deformations, combined with the relatively small sample sizes available, make statistical analysis challenging. However, the intrinsic dimensionality of brain shape variability is much lower. Extracting these intrinsic dimensions before further statistical analysis can improve the statistical power and interpretability of results.

Motivated by Bayesian reasoning, current approaches in diffeomorphic atlas building [6,12,14] are formulated as *maximum a posteriori* (MAP) optimization problems. A set of input images are registered to a template, which is simultaneously estimated with the unknown deformations in an alternating optimization strategy. In these approaches, the likelihood is defined by an image match term which is a sum squared distance function between deformed atlas and input image, and a prior on transformations that enforces smoothness. Allassonnière et al. [1] proposed a fully generative Bayesian model of elastic deformation in which estimation proceeds by marginalization over the latent image transformations.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 121–128, 2014.

Ma et al. [7] introduced a Bayesian formulation of the diffeomorphic image atlas problem by adding fixed hypertemplate information. Simpson et al. [10] inferred the level of regularization in small deformation registration by a hierarchical Bayesian model. Zhang et al. [18] develop a generative model for diffeomorphic atlas formulation and regularization parameter estimation by using a Monte Carlo Expectation Maximization (MCEM) algorithm.

Beyond estimation of an atlas, or mean image, several dimensionality reduction methods have been proposed for modeling shape variability in the diffeomorphism setting. Vaillant et al. [13] compute a PCA in the tangent space to the atlas image. Later, Qiu et al. [9] used TPCA as an empirical shape prior in diffeomorphic surface matching. Gori et al. [5] formulate a Bayesian model of shape variability using diffeomorphic matching of currents. Their model includes estimation of a covariance matrix of the deformations, from which they then extract PCA modes of shape variability. Even though these methods formulate the atlas and covariance estimation as probabilistic inference problems, the dimensionality reduction is done after the fact, i.e., as a singular value decomposition of the covariance as a second stage after the estimation step. We propose instead to treat the dimensionality reduction step as a probabilistic inference problem on discrete images, in a model called Bayesian principal geodesic analysis (BPGA), which jointly estimates the image atlas and principal geodesic modes of variation. This Bayesian formulation has two advantages. First, computing a PCA after the fact in the tangent space does not explicitly optimize the fit of the principal modes to the data (this is due to the nonlinearity of the space of diffeomorphisms), whereas we explicitly optimize this criteria intrinsically in the space of diffeomorphisms, resulting in better fits to the data. Second, by formulating dimensionality reduction as a Bayesian model, we can also infer the inherent dimensionality directly from the data.

Our work is inspired by the Bayesian PCA model introduced in Euclidean space by Bishop (BPCA) [2]. Recently, Zhang and Fletcher [17] introduced a probabilistic principal geodesic analysis (PPGA) to finite-dimensional manifolds based on PGA [3]. This work goes beyond the PPGA model by introducing the automatic dimensionality reduction, as well as extending from finite-dimensional manifolds to the infinite-dimensional case of diffeomorphic image registration. We also mention the relationship of our work to manifold learning approaches to dimensionality reduction [4]. The main advantage of the Bayesian approach we present is that it is fully *generative*, and the principal modes of variation can reconstruct shape deformation of individuals, information that is lost when mapping to a Euclidean parameter space in manifold learning. We show experimental results of principal geodesics and parameters estimated from OASIS brain dataset. To validate the advantages of our model, we reconstruct images from our estimation and compare the reconstruction errors with TPCA of diffeomorphisms and LPCA based on image intensity. Our results indicate that intrinsic modeling of the principal geodesics, estimated jointly with the image atlas, provides a better description of brain image data than computing PCA in the tangent space after atlas estimation.

2 Background

We define a generative probabilistic model for principal geodesic analysis in the setting of diffeomorphic atlas building. Before introducing our model, we first briefly review the mathematical background of diffeomorphic atlas building and its computations for geodesic shooting [11,15,16]. We use vector-valued momenta [11], which unlike scalar momenta, decouple the deformations from the atlas, leading to more efficient and stable estimation procedures.

In this framework, given input images $I_1, \ldots, I_N \in L^2(\Omega, \mathbb{R})$, a minimization problem is solved to estimate the template image and the diffeomorphic transformations between the template and each input image as

$$E(v^k, I) = \sum_{k=1}^{N} \frac{1}{2\sigma^2} \left\| I \circ (\phi_k)^{-1} - I_k \right\|^2 + \int_0^1 (L v_t^k, v_t^k)\, dt. \tag{1}$$

Here σ^2 represents noise variance, and the $v^k \in L^2([0,1], V)$ are time-varying velocity fields in a reproducing kernel Hilbert space, V, equipped with a metric, $L : V \to V^*$, a positive-definite, self-adjoint, differential operator, mapping to the dual space, V^*. The dual to the vector v^k is a momentum, $m^k \in V^*$, such that $m^k = Lv^k$ and $v^k = Km^k$. The operator K is the inverse of L. The notation (m^k, v^k) denotes the pairing of a momentum vector $m^k \in V^*$ with a tangent vector $v^k \in V$. The deformation ϕ_k is defined as the integral flow of v^k, that is, $(d/dt)\phi_k(t,x) = v^k(t, \phi_k(t,x))$. We use subscripts for the time variable, i.e., $v_t(x) = v(t,x)$, and $\phi_t(x) = \phi(t,x)$. When the energy above is minimized over the initial momenta m^k, the geodesic path ϕ_k is constructed via integration of the following EPDiff equation [8]:

$$\frac{\partial m^k}{\partial t} = -\mathrm{ad}^*_{v^k} m^k = -(Dv^k)^T m^k - Dm^k\, v^k - m^k \,\mathrm{div}(v^k), \tag{2}$$

where D denotes the Jacobian matrix, and the operator ad^* is the dual of the negative Lie bracket of vector fields, $\mathrm{ad}_v w = -[v,w] = Dvw - Dwv$.

3 Bayesian Principal Geodesic Analysis

3.1 Probability Model

We formulate the random momentum for the kth individual as $m^k = Wx^k$, where W is a matrix with q columns of principal initial momenta, $x \in \mathbb{R}^q$ is a latent variable that lies in a low-dimensional space, with $x \sim N(0, I)$. Our noise model is i.i.d. Gaussian at each image voxel, with likelihood given by

$$p(I_k \mid W, I, \sigma) = \frac{1}{(2\pi)^{M/2}\sigma^M} \exp\left(-\frac{\|I \circ (\phi_k)^{-1} - I_k\|^2}{2\sigma^2} \right), \tag{3}$$

where M is the number of voxels, and the norm inside the exponent is the $L^2(\Omega, \mathbb{R})$ norm. Note that for a continuous domain, this is not a well-defined

probability distribution due to its infinite measure on images. Therefore, we consider the input images as well as diffeomorphisms to be defined on a finite discretized grid.

The prior on W is given by the combination of a multivariate Gaussian distribution on the initial momenta m that guarantees smoothness of the geodesic shooting path, and a Gaussian distribution on W to suppress small principal initial momenta to zero. This second term is analogous to the automatic relevance determination (ARD) prior used in BPCA [2], with the difference that we use the natural Hilbert space norm for the momenta. This prior induces sparsity in the columns of W and automatically selects the dimensionality. The formulation is given by

$$p(W \mid x, \gamma) \propto \left(\prod_{i=1}^{q} \left(\frac{\gamma_i}{2\pi} \right)^{\frac{d}{2}} \right) \exp \left(-\frac{1}{2} \sum_{k=1}^{N} \|m^k\|_K^2 - \sum_{i=1}^{q} \frac{\gamma_i}{2} \|W_i\|_K^2 \right), \quad (4)$$

where i denotes the ith principal initial momentum, and γ_i is a hyperparameter which controls the precision of the corresponding W_i. Estimating γ_i induces sparsity so that if it has a large value, then the corresponding W_i will become small, and will be effectively removed in the latent space. In this work, we use a metric of the form $K = (-\alpha\Delta + \beta I)^{-2}$, where Δ is the discrete Laplacian. In this operator, α controls the smoothness of diffeomorphisms, and β is a small positive number to ensure that the K operator is nonsingular.

3.2 Inference

After defining the likelihood (3) and prior (4) in the previous section, we now arrive at the log joint posterior for the diffeomorphisms as

$$\begin{aligned} \log \prod_{k=1}^{N} p\left(W \mid I_k; I, \sigma^2, \gamma\right) \propto & -\frac{1}{2} \sum_{k=1}^{N} \|m^k\|_K^2 - \frac{1}{2\sigma^2} \sum_{k=1}^{N} \|I \circ (\phi_k)^{-1} - I_k\|^2 \\ & - \frac{MN}{2} \log \sigma - \sum_{i=1}^{q} \left[\frac{\gamma_i}{2} \|W_i\|_K^2 - \log \gamma_i \right] - \frac{1}{2} \|x^k\|^2. \end{aligned} \quad (5)$$

We use MAP estimation to determine the model parameters $\theta = I, \sigma$. In order to treat the x^k as latent random variables with log posterior given by (5), we would ideally integrate out the latent variables, which is intractable in closed form. Instead, we use a mode approximation to the posterior distribution. Next, we introduce a gradient ascent scheme to estimate W, x^k, $\theta = (I, \sigma)$ and γ simultaneously.

Gradient Ascent for W, x^k: We need to compute the gradient with respect to initial momentum m^k of the diffeomorphic image matching problem in (1), and then apply the chain rule to obtain the gradient term w.r.t. W and x^k. Following the optimal control theory approach in [15], we add Lagrange multipliers to

constrain the kth diffeomorphism $\phi_k(t)$ to be a geodesic path. The following equations are equivalent for the geodesic paths of each of the subjects, so for notational simplicity, we drop the subject index k from the notation momentarily. This is done by introducing time-dependent adjoint variables, $\hat{m}, \hat{I}$ and $\hat{v}$, and writing the augmented energy

$$\tilde{E}(m) = E(Km, I, I_k) + \int_0^1 \left[\langle \hat{m}, \dot{m} + \mathrm{ad}_v^* m \rangle + \langle \hat{I}, \dot{I} + \nabla I \cdot v \rangle + \langle \hat{v}, m - Lv \rangle \right] dt,$$

where E is the diffeomorphic image matching energy from (1), and the other terms correspond to Lagrange multipliers enforcing: a) the geodesic constraint, which comes from the EPDiff equation (2), b) the image transport equation, $\dot{I} = -\nabla I \cdot v$, and c) the constraint that $m = Lv$, respectively.

The optimality conditions for m, I, v are given by the following time-dependent system of ODEs, termed the *adjoint equations*:

$$-\dot{\hat{m}} + \mathrm{ad}_v \hat{m} + \hat{v} = 0, \qquad -\dot{\hat{I}} - \nabla \cdot (\hat{I} v) = 0, \qquad -\mathrm{ad}_{\hat{m}}^* m + \hat{I} \nabla I - L\hat{v} = 0,$$

subject to initial conditions $\hat{m}(1) = 0, \hat{I}(1) = \frac{1}{\sigma^2}(I(1) - I_k)$. Finally, after integrating these adjoint equations backwards in time to $t = 0$, the gradient of $\tilde{E}$ with respect to the initial momentum is $\nabla_m \tilde{E} = Km - \hat{m}$.

Applying the chain rule, the gradient term of (5) for updating W is

$$\nabla_W \tilde{E} = -\sum_{k=1}^{N} (Km^k - \hat{m^k})[x^k]^T - KW\gamma,$$

where γ is a diagonal matrix with diagonal element γ_i. The gradient with respect to x^k is

$$\nabla_{x^k} \tilde{E} = -W^T (Km^k - \hat{m^k}) - x^k.$$

Closed-Form Solution for θ, γ: We now derive the maximization for updating the parameters θ. This turns out to be a closed-form update for the atlas I, noise variance σ^2, and dimensionality control parameter γ. For updating the atlas image I, we set the derivative of the log posterior with respect to I to zero. The solution for I, σ^2 gives an update

$$I = \frac{\sum_{k=1}^{N} I_k \circ \phi_k |D\phi_k|}{\sum_{k=1}^{N} |D\phi_k|}, \qquad \sigma^2 = \frac{1}{MN} \sum_{k=1}^{N} \| I \circ (\phi_k)^{-1} - I_k \|^2.$$

We use the similar approximation on ARD prior in BPCA [2] to get a closed-form update for γ_i, as $\gamma_i = q / \|W_i\|_K^2$.

4 Results

4.1 OASIS Brain Dataset

To demonstrate the effectiveness of our proposed model and MAP estimation, we applied our BPGA model to a set of axial slices of brain magnetic resonance

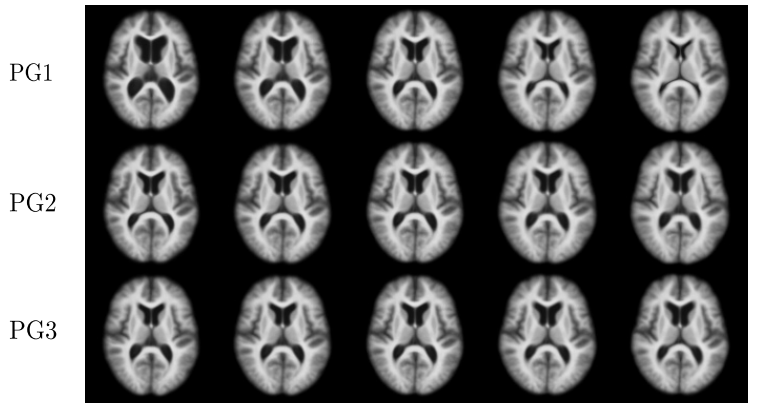
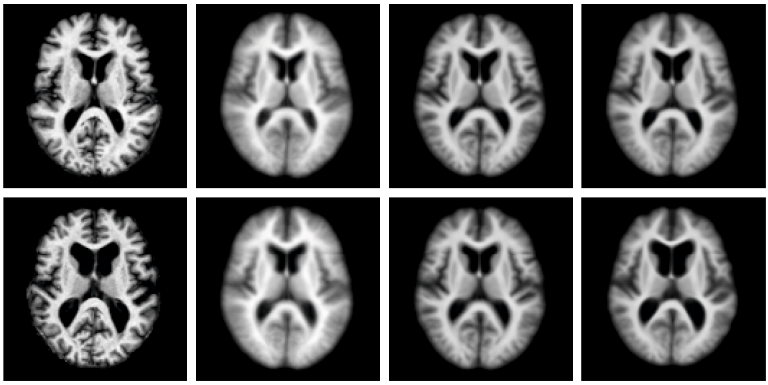


Fig. 1. Top to bottom: shooting atlas by the first, second and third principal modes. Left to right: BPGA model of image variation evaluated at $a_i = -3, -1.5, 0, 1.5, 3$.

images (MRI) from the OASIS brain database. The data consists of MRI from 40 healthy subjects between the age of 60 to 95. The MRI have resolution $108 \times 128 \times 128$ and are skull-stripped, intensity normalized, and co-registered with rigid transforms. We use $\alpha = 0.8$, $\beta = 0.4$ estimated using the procedure in [18] with 15 time-steps in geodesic shooting, and initialize the template I as the average of image intensities, while W as the matrix of principal components from TPCA.

The proposed BPGA model automatically determined that the latent dimensionality of the data was three. Figure 1 displays the automatic estimated modes, $i = 1, 2, 3$, of the brain MRI variation. We forward shoot the constructed atlas, I, by the estimated principal momentum $a_i W_i$ along geodesics. For visualization purpose, here we demonstrate the brain variation from the atlas by $a_i = -3, -1.5, 0, 1.5, 3$. The first mode of variation clearly shows that ventricle size change is a dominant source variability in brain shape. The algorithm also jointly estimated the image noise standard deviation parameter as $\sigma = 0.04$.

Image Registration Accuracy. We validated the image registration accuracy of our BPGA model. After estimating the principal initial momenta and parameters from the training subjects above, we used these estimates to reconstruct another 20 testing subjects from the same OASIS database that were not included in the training. We then measured the discrepancy between the reconstructed images and the testing images. Note that our reconstruction only used the first three principal modes, which were automatically selected by our algorithm. We also compared our model with LPCA and TPCA, also using the first three dimensions. Examples of the reconstructed images from these models are shown in Figure 2. Table 1 shows the comparison of the registration accuracy as measured by the average and standard deviation of the mean squared error (MSE). It indicates that our model outperforms both LPCA and TPCA in the diffeomorphic setting.



(a) Observed (b) LPCA (c) TPCA (d) BPGA

Fig. 2. Left to right: original data, reconstruction by LPCA, TPCA, and BPGA

Table 1. Comparison of mean squared reconstruction error between LPCA, TPCA and BPGA models. Average and standard deviation over 20 test images.

	LPCA	TPCA	BPGA
Average MSE	2.8×10^{-2}	1.6×10^{-2}	1.1×10^{-2}
Std of MSE	7.5×10^{-3}	2.3×10^{-3}	2.0×10^{-3}

5 Conclusion

We presented a generative Bayesian model of principal geodesic analysis in diffeomorphic image registration. Our method is the first probabilistic model for automatic dimensionality reduction for diffeomorphisms. We developed an inference strategy based on MAP to estimate parameters, including the noise variance and image atlas, simultaneously. The estimated low-dimensional latent variables provide a compact representation of the anatomical variability in a large image database, and they can be used for further statistical analysis of anatomical shape in clinical studies. Reducing the dimensionality to the inherent modes of shape variability has the potential to improve hypothesis testing, classification, mixture models, etc.

Acknowledgments. This work was supported by NIH Grant 5R01EB007688 and NSF CAREER Grant 1054057.

References

1. Allassonnière, S., Kuhn, E.: Stochastic algorithm for parameter estimation for dense deformable template mixture model. ESAIM-PS 14, 382–408 (2010)
2. Bishop, C.M.: Bayesian pca. In: Advances in Neural Information Processing Systems, pp. 382–388 (1999)

3. Fletcher, P.T., Lu, C., Joshi, S.: Statistics of shape via principal geodesic analysis on Lie groups. In: Computer Vision and Pattern Recognition, pp. 95–101 (2003)
4. Gerber, S., Tasdizen, T., Fletcher, P.T., Joshi, S., Whitaker, R.: Manifold modeling for brain population analysis. Medical Image Analysis 14(5), 643–653 (2010)
5. Gori, P., et al.: Bayesian atlas estimation for the variability analysis of shape complexes. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part I. LNCS, vol. 8149, pp. 267–274. Springer, Heidelberg (2013)
6. Joshi, S., Davis, B., Jomier, M., Gerig, G.: Unbiased diffeomorphic atlas construction for computational anatomy. NeuroImage 23(suppl. 1), 151–160 (2004)
7. Ma, J., Miller, M.I., Trouvé, A., Younes, L.: Bayesian template estimation in computational anatomy. NeuroImage 42, 252–261 (2008)
8. Miller, M.I., Trouvé, A., Younes, L.: Geodesic shooting for computational anatomy. Journal of Mathematical Imaging and Vision 24(2), 209–228 (2006)
9. Qiu, A., Younes, L., Miller, M.I.: Principal component based diffeomorphic surface mapping. IEEE Transactions on Medical Imaging 31(2), 302–311 (2012)
10. Simpson, I.J.A., Schnabel, J.A., Groves, A.R., Andersson, J.L.R., Woolrich, M.W.: Probabilistic inference of regularisation in non-rigid registration. NeuroImage 59, 2438–2451 (2012)
11. Singh, N., Hinkle, J., Joshi, S., Fletcher, P.T.: A vector momenta formulation of diffeomorphisms for improved geodesic regression and atlas construction. In: International Symposium on Biomedical Imaging, pp. 1219–1222 (2013)
12. Twining, C.J., Cootes, T.F., Marsland, S., Petrovic, V., Schestowitz, R., Taylor, C.J.: A unified information-theoretic approach to groupwise non-rigid registration and model building. In: Christensen, G.E., Sonka, M. (eds.) IPMI 2005. LNCS, vol. 3565, pp. 1–14. Springer, Heidelberg (2005)
13. Vaillant, M., Miller, M.I., Younes, L., Trouvé, A.: Statistics on diffeomorphisms via tangent space representations. NeuroImage 23, S161–S169 (2004)
14. Vialard, F.X., Risser, L., Holm, D., Rueckert, D.: Diffeomorphic atlas estimation using Kärcher mean and geodesic shooting on volumetric images. In: MIUA (2011)
15. Vialard, F.X., Risser, L., Rueckert, D., Cotter, C.J.: Diffeomorphic 3d image registration via geodesic shooting using an efficient adjoint calculation. International Journal of Computer Vision, 229–241 (2012)
16. Younes, L., Arrate, F., Miller, M.: Evolutions equations in computational anatomy. NeuroImage, 40–50 (2009)
17. Zhang, M., Fletcher, P.T.: Probabilistic principal geodesic analysis. In: Advances in Neural Information Processing Systems, pp. 1178–1186 (2013)
18. Zhang, M., Singh, N., Fletcher, P.T.: Bayesian estimation of regularization and atlas building in diffeomorphic image registration. In: Gee, J.C., Joshi, S., Pohl, K.M., Wells, W.M., Zöllei, L. (eds.) IPMI 2013. LNCS, vol. 7917, pp. 37–48. Springer, Heidelberg (2013)

New Partial Volume Estimation Methods for MRI MP2RAGE

Quentin Duché[1,2,3,5], Parnesh Raniga[5,6], Gary F. Egan[6], Oscar Acosta[1,2], Giulio Gambarota[1,2,3], Olivier Salvado[5], and Hervé Saint-Jalmes[1,2,3,4]

[1] Université de Rennes 1, LTSI - Rennes, F-35000, France
[2] INSERM, U1099 - Rennes, F-35000, France
[3] PRISM - Biosit, CNRS UMS 3480 - Biogenouest - Rennes, F-35000, France
[4] CRLCC, Centre Eugène Marquis - Rennes, F-35000, France
[5] The Australian E-Health Research Centre, CSIRO Preventative Health Flagship, CSIRO Computational Informatics, Herston, QLD, Australia
[6] Monash Biomedical Imaging, Melbourne, VIC, Australia
quentin.@univ-rennes1.fr

Abstract. Magnetic resonance imaging (MRI) is commonly used as a medical diagnosis tool, especially for brain applications. Some limitations affecting image quality include receive field (RF) inhomogeneity and partial volume (PV) effects which arise when a voxel contains two different tissues, introducing blurring. The novel Magnetization-Prepared 2 Rapid Acquisition Gradient Echoes (MP2RAGE) provides an image robust to RF inhomogeneity. However, PV effects are still an issue for automated brain quantification. PV estimation methods have been proposed based on computing the proportion of one tissue with respect to the other using linear interpolation of pure tissue intensity means. We demonstrated that this linear model introduces bias when used with MP2RAGE and we propose two novel solutions. The PV estimation methods were tested on 4 MP2RAGE data sets.

Keywords: MP2RAGE, Partial Volume Estimation, Bi-exponential model.

1 Introduction

Magnetic resonance imaging (MRI) is a commonly used modality for brain diagnosis and many morphometric methods have been developed to estimate brain atrophy [1,2,3]. However MRI has some limitations, which may affect the performance of several image processing steps and may hamper automated structural quantification if not taken into account. Among them, noise, the receive field (RF) inhomogeneity and partial volume (PV) effects.

The novel Magnetization-Prepared 2 Rapid Acquisition Gradient Echoes (MP2RAGE) sequence [4] has good signal-to-noise and contrast properties and is therefore an excellent candidate for image processing methods. The sequence also tackles the inhomogeneity of the signal across the scanned volume with a double acquisition approach. Two co-registered images are obtained and both are identically biased. A composite image is computed free of any RF inhomogeneity. MP2RAGE was also designed to maximise contrast-to-noise ratio per unit of time between brain tissues to

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 129–136, 2014.

facilitate segmenting the brain in main tissues: gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF).

However, PV effects remain an issue in MP2RAGE. They occur when two different tissues, having different magnetic properties, contribute to the signal of a single voxel. PV estimation (PVE) consists in assigning a fractional content, *i.e.* a proportion, to each of the tissues composing a voxel labeled as a PV voxel. PVE has been shown to be useful in cortical thickness estimation [5,6,7] as the cerebral cortex, the GM, is surrounded by two different tissues: WM and CSF. Thus, the cortex is subject to two types of PV effects at its two interfaces: GM/WM and GM/CSF. Additionally, cortical thickness is of the same order of magnitude as the image resolution (typically a few mm). Its size and its convoluted structure make the cortex very sensitive to PV effects.

As cortical thickness reduction has been shown to be a good biomarker for many neurodegenerative diseases such as Alzheimer's [1], we focus this paper at estimating PV effects using MP2RAGE.

Previous works [8,9] rely on the same PV model [7] to estimate fractional contents and calculate PV maps. This model has not been validated on the particular MP2RAGE sequence yet. In this work, we evaluate the commonly used PV model on the composite image computed with the two acquisitions in MP2RAGE.

2 MP2RAGE

MP2RAGE is a recent sequence based on the popular MPRAGE sequence [10]. It starts with a magnetization preparation followed by two gradient echo blocks providing two co-registered and differently contrasted images S_1 and S_2 (Fig. 1(a) and (b)). MP2RAGE has the advantage of being robust to the RF inhomogeneities as a composite image U (Fig. 1(c)) is computed inline with the two images in a way that cancels the RF inhomogeneity:

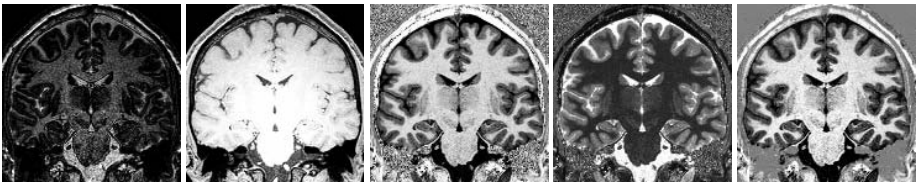
$$U = \frac{Real(S_1^* S_2)}{|S_1|^2 + |S_2|^2} \tag{1}$$

where the symbol * stands for the complex conjugate, more details regarding this equation can be found in [4]. Eq. (1) constrains the possible values in U between -0.5 and 0.5. U is not linear with respect to S_1 and S_2. This sequence also has the advantage of producing a high resolution T1 map (Fig. 1(d)). For tissues with a long longitudinal relaxation time T1, the short first inversion time in MP2RAGE results in negative signals. The sign information associated with S_1 was estimated by assuming that S_2 has positive signals due to the late second inversion time for the brain tissues considered, and therefore the sign of U is a good estimator for the sign associated with S_1. This allows using the entire dynamic range of S_1 in a new image called $S_1^{\pm}$ (Fig. 1(e)).

$$S_1^{\pm} = \frac{U(S_1^2 + S_2^2)}{S_2} \tag{2}$$

3 Methods

PV classes are often modeled as a linear mixture of two normal distributions modeling two pure tissues [11,7,5]. Under this assumption, the maximum likelihood estimation



(a) S_1 (b) S_2 (c) U (d) $T1$ map (e) $S_1^{\pm}$

Fig. 1. MP2RAGE images shown in a coronal plane. The sequence measures S_1 and S_2. U is computed inline with the two acquisitions with Eq. (1). The $T1$ map is also estimated inline. Our preprocessing includes reconstructing $S_1^{\pm}$ with the sign information contained in U.

of the mixture coefficient is a linear interpolation of tissue intensity means. In this section, the traditional linear model for PV effects will be investigated for the MP2RAGE sequence. The purpose of this paper is PVE only, in other words, the fractional content calculation. We assume that the brain tissues have already been segmented from the composite image U into GM, WM and CSF using an established and well validated method [11]. In this paper, we want to compare the GM fractional content estimated with three PVE methods at the GM boundaries. The explanations on the PVE models are concerned with a GM/WM voxel for the sake of clarity, but a similar reasoning can be applied to a GM/CSF voxel. The unknown GM fractional content is called $\alpha \in [0, 1]$.

3.1 Linear Interpolation of Intensity Means (LIMe)

In the majority of previous works on PVE, regardless of the sequence, the signal s_{gw} of a voxel composed of GM and WM is modeled as a linear combination of intensity means (μ_g and μ_w) of pure tissues

$$s_{gw} = \alpha\mu_g + (1 - \alpha)\mu_w \tag{3}$$

The model is parameterized by pure tissue intensity means. The fractional content calculation is done by interpolating the signal s_{gw} as following. f restricts the value of α in $[0, 1]$:

$$\alpha = f(\frac{\mu_w - s_{gw}}{\mu_w - \mu_g}) \tag{4}$$

The linear PV model could be independently applied to S_2 or $S_1^{\pm}$ but RF insensitivity and the optimized contrasts between cerebral tissues obtained in U would not be exploited. Given that the composite image is not linearly obtained, the well-known linear PV model (Eq.(3)) introduces errors. Assuming that partial voluming is linear in α in images $S_1^{\pm}$ and S_2, the linear model could be applied independently to both images. We call g_1 and g_2 (respectively w_1 and w_2) the intensity means of pure GM (respectively WM) in $S_1^{\pm}$ and S_2. Thus, given Eq. (1) and neglecting the noise, the GM/WM PV signal U_{gw} obtained in U can be expressed as:

$$\begin{cases} s_{1gw} = \alpha g_1 + (1 - \alpha)w_1 \\ s_{2gw} = \alpha g_2 + (1 - \alpha)w_2 \end{cases} \Rightarrow U_{gw} = \frac{s_{1gw}s_{2gw}}{s_{1gw}^2 + s_{2gw}^2} \tag{5}$$

$$U_{gw} = \frac{\boldsymbol{\alpha^2}(g_1g_2 + w_1w_2 - g_1w_2 - w_1g_2) + \boldsymbol{\alpha}(g_1w_2 + w_1g_2 - 2w_1w_2) + w_1w_2}{\boldsymbol{\alpha^2}((g_1 - w_1)^2 + (g_2 - w_2)^2) + 2\boldsymbol{\alpha}(g_1w_1 + g_2w_2 - w_1^2 - w_2^2) + w_1^2 + w_2^2} \tag{6}$$

From Eq. (6), it is clear that partial voluming in U is not linear but quadratic in α. Assuming a linear PV model for $S_1^{\pm}$ and S_2 results in a non-linear model for U. We propose in the next sections new models that reduce this error.

3.2 Quadratic Interpolation of Intensity Means (QIMe)

The first solution that we propose to address PVE in U is an extension of LIMe, using a Quadratic Interpolation of Intensity Means (QIMe). As the PV signal in U appears to be quadratic, finding α is equivalent to solving a second order equation with the following reformulation of Eq.(6):

$$\begin{aligned} U_{gw} = \tfrac{N_{gw}(\alpha)}{D_{gw}(\alpha)} &\Leftrightarrow N_{gw}(\alpha) = U_{gw}D_{gw}(\alpha) \\ &\Leftrightarrow N_{gw}(\alpha) - U_{gw}D_{gw}(\alpha) = 0 \\ &\Leftrightarrow P_{gw}(\alpha) = 0 \end{aligned}$$

Finding the fractional content is equivalent to finding the roots of a second order polynom for every PV voxel. This polynom is parameterized by the signal U_{gw} and the intensity means of pure tissue in $S_1^{\pm}$ and S_2. When the discriminant of P_{gw} (Δ) is negative, there are no computable solutions so α is set to the closest tissue in terms of intensity in U (0 for WM, 1 for GM). When $\Delta > 0$, the closer root to the LIMe solution is chosen as the evolution of α as a function of a PV signal appears to be almost linear.

3.3 Bi-Exponential Model (BiExp)

Duché et al. [12] proposed a bi-exponential model (BiExp) to estimate PV from MP2RAGE. In BiExp, the parameters contributing to the signal are expressed: the tissue properties and the sequence parameters. Hence, the signal measured in a voxel is weighted by the longitudinal magnetization of the protons population M_0. Consequently, the two PV signals in $S_1^{\pm}$ and S_2 are defined as a linear combination of two pure signals:

$$\begin{cases} s_{1gw} &= M_{0g}s_1(T1_g) + M_{0w}s_1(T1_w) &= M_{0g}s_{1g} + M_{0w}s_{1w} \\ s_{2gw} &= M_{0g}s_2(T1_g) + M_{0w}s_2(T1_w) &= M_{0g}s_{2g} + M_{0w}s_{2w} \end{cases} \tag{7}$$

where $T1_g$ and $T1_w$ are the $T1$ values of pure GM and WM. They are estimated in the T1 map produced by MP2RAGE. The signals s_1 and s_2 have been described by Marques *et al.* [4], they are functions of the many sequence parameters and magnetic properties of the tissues. The assumption of this model is the uniqueness of T1 value per tissue. This simplifies the system as the signals s_1 and s_2 can then be computed for particular T1 values, resulting in the estimation of the constant $s_{1g}, s_{2g}, s_{1w}, s_{2w}$. They represent the pure GM and WM signals in $S_1^{\pm}$ and S_2 for $M_0 = 1$.

This voxel-wise linear system can be solved for (M_{0g}, M_{0w}) which are the amounts of respective pure tissues in the voxel, they represent the same physical information in

both co-registered MP2RAGE images. The fractional content of GM is calculated as $\alpha = \frac{M_{0g}}{M_{0g}+M_{0w}}$. This model is parameterized by the T1 values of pure tissues. T2 has a limited impact on α and the proton density values of the tissues are taken from the literature [13,14].

Table 1. Summary of the three presented PV estimation methods (first column). The second column names the required parameters for the method and recalls the image(s) they are extracted from. The last column contains the image(s) in which the PV estimation is done.

PVE Method	Parameters (extracted from)	PV estimated with
LIMe	μ_g, μ_w, μ_c (U)	U
QIMe	$g_{1,2}, w_{1,2}, c_{1,2}$ ($S_1^{\pm}, S_2$)	U
BiExp	$T1_g, T1_w, T1_c$ ($T1$ map)	$S_1^{\pm}, S_2$

4 Experiments

4.1 Simulations

The three tissues were simulated with the T1 measured in the experimental data. Each GM interface was discretized with intermediate PV values where the signal was modeled as a linear combination of two pure tissue signals. The noiseless two echoes and composite signals were simulated. The three methods were applied to estimate the fractional content α. PVE by the various methods was expressed as a function f_s of the ground truth (GT) α.

4.2 Experimental Data

Two healthy volunteers were scanned twice in a 3T Siemens Scanner with a 20-channel head coil. A 3D isotropic (1mm^3) MP2RAGE protocol was used. Each MP2RAGE data underwent identical pre-processing that included brain extraction and automated segmentation of GM, WM and CSF. These masks were eroded to estimate parameters for the three PVE methods, the erosion allows to avoid a large number of voxels subject to PV effects at the boundaries. The parameters estimations were done in the same regions for the three methods. GM PV maps were calculated with the three methods presented in section 3. Boundary masks (GM/WM and GM/CSF) were extracted by taking the intersection of the dilated segmentations. This ensures to define regions in which a majority of the voxels are subject to PV effects. In these boundaries voxels, the three PV methods were compared with the same population of voxels.

Experimental results from the four scans were gathered for the analysis. The GM voxels were separated in two classes resulting in about 1 million GM/WM voxels and 1.5 million GM/CSF voxels. Fractional content estimates from two different PV methods were plotted and treated as a joint probability distribution. These 2D histograms were integrated to get an average function. These experimental functions f_e were compared with the function f_s obtained in the simulation.

5 Results and Discussion

Results from the simulations and experimental data are summarized in Fig. 2. The first row shows the MP2RAGE U image and the extracted GM boundaries. In the second

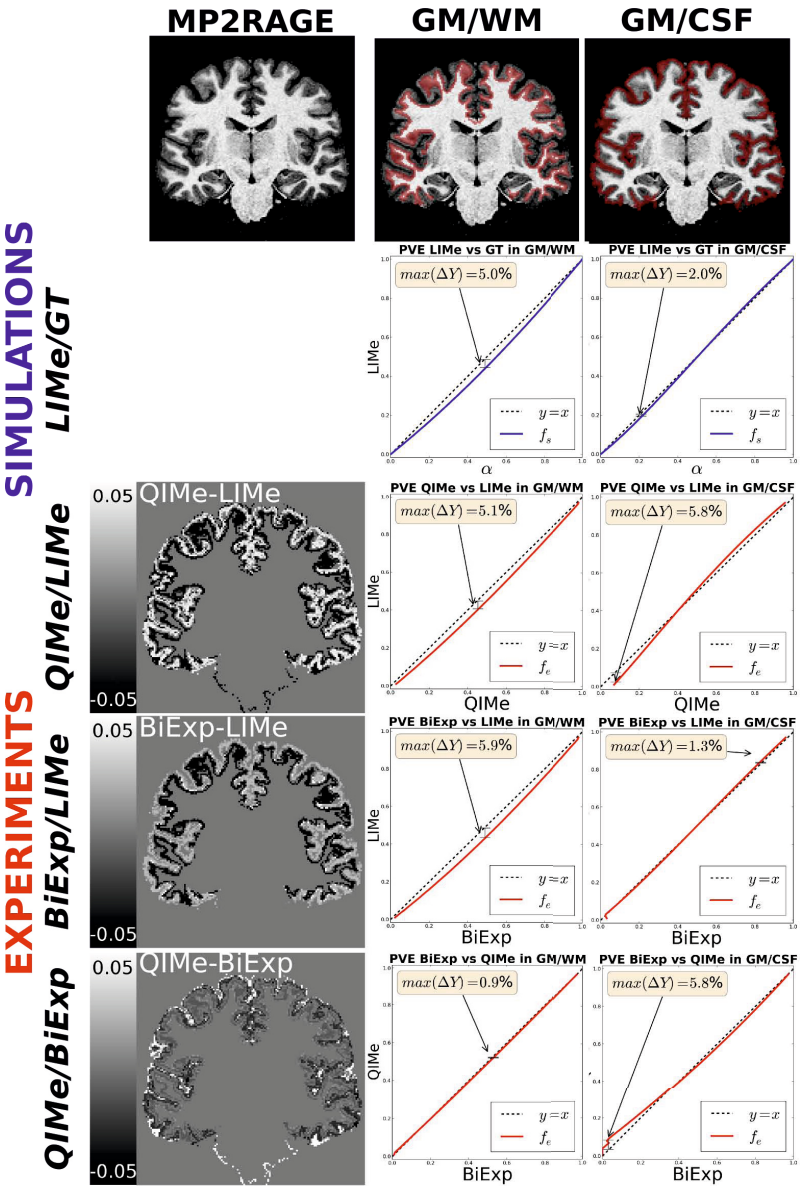


Fig. 2. Summary of the results. The first row shows the input and the two GM interfaces of interest where fractional content estimates are extracted from. The second row shows the results of the simulations for LIMe. For the experimental results, every row is the comparison of two PVE methods, the image is the difference of the PV maps.

row, f_s, expressing the LIMe PV GM fractional content estimate is plotted as a function of the GT α, is plotted in blue for the two boundaries. These graphs confirm the systematic errors made on the fractional content estimate with the linear PV model. It also suggests that the theoretical maximal cumulated error on both boundaries can go up to 7% of the voxel resolution used. For a 1mm^3 resolution and a cortical thickness of 3mm, this represents a maximal error of 2.33%.

In the "Experiments" part of Fig. 2, each row represents the comparison of two PVE methods as indicated on the left side by a vertical text. The image is the difference between the computed GM PV maps. The graphs exhibit the plot of f_e in red, expressing the average fractional content estimate of the first method as a function of the fractional content estimated with the second method for a large population of PV voxels.

For the GM/WM PV voxels, the experimental functions f_e for QIMe and BiExp are very similar to the noiseless simulated f_s confirming the error that we expected from using the linear model. Our results suggest that QIMe and BiExp are good PV models for MP2RAGE. When comparing QIMe to BiExp (last row), almost no difference was observed on the GM/WM boundary.

At the GM/CSF interface, the results obtained with QIMe are less consistent with the simulations when $\alpha \to 0$, *i.e.* when the voxel tends to be pure CSF. We hypothesized that this could be due to the low CSF SNR. BiExp seems more consistent with the expected behaviour of a good PV model.

QIMe and BiExp use the two echoes and are not subject to RF inhomogeneity. QIMe has the advantage of being self-contained, there are no assumptions on the T2 nor the proton density of the tissues. BiExp has the advantage of taking into account MR acquisition parameters and therefore could be extended to incorporate a model of a transmit field (TF) inhomogeneity map. The knowledge of the actual flip angle induced to the protons could be incorporated voxel-wise in the model as the signal equations are fully expressed.

The unique processing of the information contained in U (LIMe) results in underestimating the GM proportion in PV voxels. This undererestimation is systematic in the GM/WM interface. These results may explain the systematic measurement of a thinner cortex with MP2RAGE compared to MEMPRAGE found in [15]. Our work may be an answer to the missing tailored tissue segmentation method needed by MP2RAGE as Fujimoto *et al.* pointed out.

6 Conclusion

We investigated the well known problem of PVE with the novel MP2RAGE sequence. The well established linear model for PVE is prone to errors on both interfaces surrounding cortical GM. Our experiments suggest that PV can not be correctly estimated with the unique analysis of combined image U, the information contained in the two images $S_1^{\pm}$ and S_2 must be exploited. We proposed two solutions which led to similar results. Both methods provide a way forward to improve the accuracy of cortical surface reconstruction with MP2RAGE. Future work will include measuring the impact of the PVE error in GM with LIMe on cortical thickness estimation.

Acknowledgement. This work is partially funded by the "Région Bretagne".

References

1. Acosta, O., et al.: Cortical surface mapping using topology correction, partial flattening and 3D shape context-based non-rigid registration for use in quantifying atrophy in Alzheimer's disease. Journal of Neuroscience Methods (2011)
2. Rueda, A., et al.: Topology-corrected segmentation and local intensity estimates for improved partial volume classification of brain cortex in MRI. Journal of Neuroscience Methods 188, 305–315 (2010)
3. Doré, V., et al.: Cross-sectional and longitudinal analysis of the relationship between Aβ deposition, cortical thickness, and memory in cognitively unimpaired individuals and in alzheimer disease. JAMA Neurology 70(7), 903–911 (2013)
4. Marques, J., et al.: MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. Neuroimage 49(2), 1271–1281 (2010)
5. Ballester, A.M., et al.: Segmentation and measurement of brain structures in MRI including confidence bounds. Convergence 4, 189–200 (2000)
6. Leemput, K.V., et al.: A Unifying Framework for Partial Volume Segmentation of Brain MR Images 18(10), 897–908 (October 1999)
7. Shattuck, D.W., et al.: Magnetic resonance image tissue classification using a partial volume model. NeuroImage M, 856–876 (2001)
8. Manjón, J., et al.: Improved estimates of partial volume coefficients from noisy brain MRI using spatial context. Neuroimage 53(2), 480–490 (2010)
9. Tohka, J., et al.: Fast and robust parameter estimation for statistical partial volume model in MRI. NeuroImage 23, 84–97 (2004)
10. Mugler, J.P., Brookeman, J.R.: Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE). Magnetic Resonance in Medicine 15(1), 152–157 (1990)
11. Acosta, O., et al.: Automated voxel-based 3D cortical thickness measurement in a combined Lagrangian-Eulerian PDE approach using partial volume maps. Medical Image Analysis 13(5), 730–743 (2009)
12. Duché, Q., Acosta, O., Gambarota, G., Merlet, I., Salvado, O., Saint-Jalmes, H.: Bi-exponential magnetic resonance signal model for partial volume computation. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part I. LNCS, vol. 7510, pp. 231–238. Springer, Heidelberg (2012)
13. Whittall, K.P., et al.: In vivo measurement of T2 distributions and water contents in normal human brain. Magnetic Resonance in Medicine 37(1), 34–43 (1997)
14. Rooney, W.D., et al.: Magnetic field and tissue dependencies of human brain longitudinal ^1H_2O relaxation in vivo 318, 308–318 (2007)
15. Fujimoto, K., et al.: Quantitative comparison of cortical surface reconstructions from MP2RAGE and multi-echo MPRAGE data at 3 and 7T. NeuroImage (2013)

Single-Subject Structural Networks with Closed-Form Rotation Invariant Matching Improve Power in Developmental Studies of the Cortex

Benjamin M. Kandel[1], Danny JJ Wang[2], James C. Gee[1], and Brian B. Avants[1]

[1] Penn Image Computing and Science Laboratory, University of Pennsylvania, Philadelphia, Pennsylvania
[2] Department of Neurology, University of California, Los Angeles, California

Abstract. Although much attention has recently been focused on single-subject functional networks, using methods such as resting-state functional MRI, methods for constructing single-subject structural networks are in their infancy. Single-subject cortical networks aim to describe the self-similarity across the cortical structure, possibly signifying convergent developmental pathways. Previous methods for constructing single-subject cortical networks have used patch-based correlations and distance metrics based on curvature and thickness. We present here a method for constructing similarity-based cortical structural networks that utilizes a rotation-invariant representation of structure. The resulting graph metrics are closely linked to age and indicate an increasing degree of closeness throughout development in nearly all brain regions, perhaps corresponding to a more regular structure as the brain matures. The derived graph metrics demonstrate a four-fold increase in power for detecting age as compared to cortical thickness. This proof of concept study indicates that the proposed metric may be useful in identifying biologically relevant cortical patterns.

1 Introduction

Brain connectivity has emerged as a dominant trend in recent neuroimaging research. Connectivity can be measured by correlation of function [6], diffusion-based structural connections [7], and covariation of cortical structure across populations [1]. Covariance patterns of cortical structure often recapitulate functional connectivity patterns [11], although this analysis is complicated in part because while functional networks can be derived on a per-subject basis, structural networks are commonly derived on a group basis, making statistical analysis of structural networks challenging [3]. Therefore, although [11] showed that ICA components of cortical covariance are similar to ICA components of rs-fMRI networks, statistical analysis on a per-subject basis is not straightforward.

As a result of these difficulties with group-wise structural networks, several groups have begun to pursue single-subject cortical networks. These methods

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 137–144, 2014.

have coalesced around two camps: Those that generate a network based on the difference between some derived scalar metrics from cortical morphology, such as cortical thickness or curvature, and those that use a patch-based correlation between two different voxels. In the first camp, [10] used a combination of cortical thickness and curvature-based metrics to construct networks. Similarly, Dai [5] used differences between regional cortical thickness measurements to create cortical networks. In the second camp, Tijms et al. [12] construct similarity networks based on the correlation between patches centered around different voxels.

Previous methods for constructing single-subject cortical networks suffer from several drawbacks. Fundamentally, cortical thickness, although an important measure of cortical structure, does not capture all the information of the surrounding neighborhood of a voxel; similarities in cortical thickness do not necessarily imply similarities in structure as a whole. Furthermore, when combining separate scalar values, such as thickness and one or several curvature measurements, how to combine the features into a meaningful distance metric is not at all straightforward, and previous methods have constructed complex and highly specialized models for individual diseases [10]. Examining correlations of patches centered around given voxels is an intuitive and straightforward approach; the only parameter to choose is the patch size, which can be chosen based on principled methods.

However, the method by Tijms for computing correlations between voxels has some technical drawbacks. To obtain rotation invariance, the method rotated cubes in increments of 45 degrees to obtain maximal correlation with the test patch. This method suffers from several flaws: First, the choice of 45 degrees is arbitrary; there is no fundamental reason to only rotate patches in increments of 45 degrees. Second, it appears that [12] "rotates" cubes by permuting the entries in the patch. This will lead to distortion in the shape, as the distance from voxel in the center of a square to the voxel immediately above it is a factor of $\sqrt{2}$ smaller than the distance between the center of the square to the corner. Third, the operation is only defined for the somewhat arbitrary shape of cubes of size $3\times3\times3$ voxels. Finally, even if the rotation were defined for arbitrary angles, an exhaustive search in three dimensions as [12] does would be computationally infeasible.

As opposed to the previously proposed approaches, we propose a closed-form, truly rotation-invariant approach to computing structural similarity across brain regions. Reorienting two images so that their orientations match is a well-posed problem that has known solutions. Leveraging these methods, we construct a rotation-invariant representation of the patches surrounding individual voxels. The correlation between different patches for different voxels give the adjacency weights in the graph. An overview is shown in Figure 1.

We apply the method to tracking the network dynamics of cortical structure in a pediatric dataset. We find that network measurements increase the power of detecting age changes by a factor of four as compared to using cortical thickness. In sum, our contributions are: 1) Method for constructing rotation-invariant structural similarity metrics; 2) Method for combining these sensibly

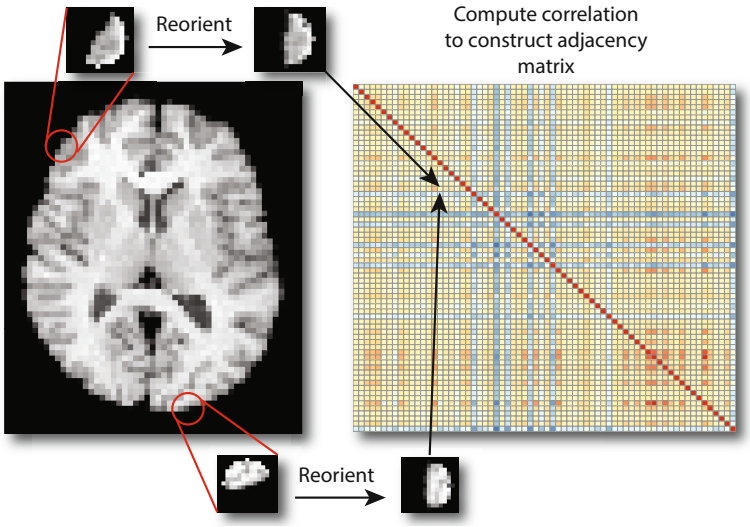


Fig. 1. Overview of adjacency matrix construction. The patches surrounding each voxel are extracted and aligned to a common reference frame. The correlation between the patches is entered into the adjacency matrix.

into a smaller-dimensional graph; 3) Demonstration that pediatric development and gender is closely correlated with node closeness, which is more predictive of age than scalar ROI values; and 4) Demonstration that the proposed method is superior to thickness distance-based cortical networks for predicting age and gender.

2 Methods

We consider a undirected graph $\mathcal{G}$ with edges $\mathcal{K}$ and nodes $\mathcal{N}$. The edges in the graph correspond to the strength of connection between different parts of the brain. Given an image $\mathcal{I}$ with J scalar-valued voxels at locations $x_j \in \mathcal{I}, j = 1, \ldots, J$, we seek a function $d : \mathcal{I}(x_i) \times \mathcal{I}(x_j) \mapsto \mathbb{R}^+$ to map from the input image space to graph edge weights. In fMRI, this function can simply be the correlation of the time-series at the different voxels, but because we consider scalar voxels, this option is not available to us. Instead, we consider the similarity between the neighborhood surrounding the voxels of interest. We denote the neighborhood of a voxel x_i as $\mathcal{N}_i = \{x_j | \ \|x_j - x_i\|_2^2 \leq r\}$, where r is the radius of the neighborhood. The edge weight between voxels x_i, x_j is then described by

$$\mathcal{K}_{i,j} = d(\mathcal{N}_i, \mathcal{N}_j). \tag{1}$$

A naïve approach to generating the function $d(x_i, x_j)$ would be to simply compute the correlation of the vector representation of $\mathcal{N}_i$ with $\mathcal{N}_j$, but this would

not account for the curved structure of the brain. Ensuring that the metric between two neighborhoods is rotation-invariant is not trivial.

Rotation-Invariant Correlation: We use a closed-form solution to align voxel neighborhoods to a canonical reference frame. Although the choice of orientation is arbitrary, we must choose one orientation as a base for reorienting all the patches. Instead of choosing one patch, which could bias our results, we first generate an $n \times m$ matrix, where each row i is the vector representation of $\mathcal{N}_i$, each of which consists of m voxels. For computational feasibility, we take a random sampling of n voxels from around the cortex. We found that no benefit was achieved by sampling more than 5000 sample voxels. The first singular vector of the sample patch matrix serves as our canonical reference frame.

Aligning the orientation of two vectors has a well-known analytical solution [9]. Aligning two images corresponds to aligning the orientations of the first eigenvector (or two eigenvectors for a 3D image) of the covariance matrix of the gradient of the image. We denote the gradient operator $g : \mathcal{N} \mapsto \mathbb{R}^{\mathrm{D}}$, where D is the number of dimensions in the image. We compute the gradient by convolving our image with the derivative of a Gaussian ($\sigma = 1$ voxel). The covariance matrix $C(\mathcal{N}_i)$ of the gradient of the neighborhood $\mathcal{N}_i$ is then given by

$$C(\mathcal{N}_i) = \sum_{x_i \in \mathcal{N}_i} g(\mathcal{I}(x_i))\, g(\mathcal{I}(x_i))^{\mathrm{T}} \in \mathbb{R}^{\mathrm{D}\times\mathrm{D}}. \tag{2}$$

To align the patches of two voxels x_i and x_j, we denote the k'th eigenvector of $C(\mathcal{N}_i)$ as w_k and the k'th eigenvector of $C(\mathcal{N}_j)$ as v_k and calculate the rotation matrix Q that best aligns them:

$$\arg\min_{Q} \sum_{k\in\{1,2\}} \|w_k - \mathbf{Q}v_k\|^2 \tag{3}$$

Denoting $B = w_k v_k^T$, we compute the singular value decomposition (SVD) of B: $B = USV^{\mathrm{T}}$. Then the analytical solution to Equation 3 is given by $Q = UMV^T$, where M = diag[1 1 det(U) det(V)]. We then rotate the voxel coordinates x_i by Q and use a linear interpolator to regenerate the neighborhood image after the rotation. Because the eigenvalues are unsigned, they can sometimes result in an alignment that is flipped by 180 degrees from the correct alignment. To eliminate this possibility, we check for a negative correlation between the sample patch and the reference patch and flip the rotation matrix if necessary. A more computationally expensive alternative is to use the Radon transform to estimate orientation [8].

Correlation Matrix Construction: In most connectome construction schemes, data is first averaged over some brain parcellation and those averaged values are then used for calculating correlations [13]. In this case, however, the average of a series of patches is ill-defined, and we found that constructing correlation matrices in this manner did not yield meaningful results. Instead, we first calculated the correlation of the vector representation of the reoriented

neighborhood of each voxel in the cortex with every other voxel in the cortex. As is standard, we constructed an $I \times I$ correlation matrix, where there are a total of I regions (nodes) for each subject. The correlation between region i and j was then calculated as the mean of the correlation of the vector representation of the reoriented neighborhood of each voxel in region i with each voxel in region j. Once the correlation matrix was constructed, normalized closeness was calculated using the `igraph` package in R [4]. For node i, closeness is defined as $\sum_{j \neq i} K_{i,j}$, with the normalization running over all nodes.

We compared our results to the method of [5], which uses the difference in cortical thickness between two regions to construct the network. Given I total regions, each with cortical thickness $t(i)$, we construct an $I \times I$ distance matrix D, where $D(i,j) = \exp - \left(\frac{(t(i)-t(j))^2}{\sigma} \right)$. We set σ to 0.015, as recommended in [5].

Clinical Pediatric Data: Our pediatric data consists of 119 subjects, with mean age 12.42, range 7.07-17.99 years, 61 females and 58 males. Magnetization-Prepared Rapid Acquisition Gradient Echo (MPRAGE) images were acquired on a Siemens Trio Tim scanner (3T) using a 3D inversion recovery sequence with TR/TE/TI = 2170/4.33/1100 ms. The resolution was 1x1x1mm^2 with a matrix size of 256x256x192. Flip angle = 7 ° and total scan time was 8:08 minutes. Image preprocessing, including bias correction, skull-stripping, segmentation, and warping of the AAL label set to the subject space was performed with ANTs [2], and the AAL label set was used for generating ROI's to construct the graphs.

Computation Considerations and Parameters: One of the advantages of a correlation-based approach to similarity evaluation is the simplicity and lack of parameters in the method. The only free parameter in this method is the patch size, which can be set based on the scale of features to be matched. Matching small patches will find similarities between small features, such as position on sulcus or gyrus, whereas matching large patches will find regional similarities. We found that downsampling images to 3mm and using a patch radius of 3 voxels was appropriate for looking at correlations between ROI's on the scale of AAL labels. Reorientation of patches takes under 20 minutes on an Intel Xeon CPU at 2.40GHz with 2 GB of memory. The code for constructing the adjacency matrix is open-source and is available at `https://github.com/bkandel/PatchAnalysis`.

3 Results

Validation of Rotation Invariance: We first checked that our output graph metrics are indeed rotation invariant. We rotated the images of ten subjects chosen at random in increments of ten degrees and constructed graphs from the rotated images. We plotted the deviation of the closeness value from the subject-wise mean closeness value vs. rotation (Figure 2 for a sample ROI).

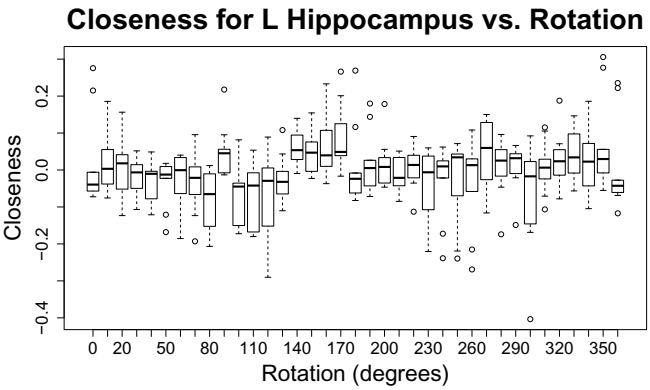


Fig. 2. Verification of rotation invariance: Difference between closeness of left hippocampus and subject-wise mean vs. rotation. The means over all rotations did not show a correlation with rotation, although there were slightly more low outliers at 90 and 270 degrees and slightly more high outliers at 0 and 180 degrees.

T-tests between the closeness values at each rotation and the mean closeness did not reveal a significant difference from the mean for any rotation (minimum FDR-corrected p-value 0.13). It may still be possible to achieve even less dependence on angle by scaling the patches to minimize the effect of outlying voxels.

Sample Subject: A thresholded correlation matrix overlaid on the MNI template brain shown in Figure 3. Cortical thickness-derived graphs tend to have many cliques, corresponding to regions with similar cortical thickness, that are not connected to each other; graphs using our method tended to have more central nodes.

Pediatric Data: In pediatric data, graph closeness was found to be highly correlated with age in most regions, whereas cortical thickness was not found to be as correlated in as many regions. To evaluate correlation of closeness with age in an ROI-wise basis, we performed an ANOVA comparing the models (in `R` notation) ROI.Closeness $\sim$ Sex + BrainVolume and ROI.Closeness $\sim$ Sex+BrainVolume+Age+Age:Sex+Age^2+Age^2:Sex, where : signifies an interaction term. Analogous ANOVA's, using the same covariates, were performed for the thickness-derived graphs and cortical thickness. Patch closeness was found to be significantly correlated with age (after FDR correction) in 68 out of 68 cortical regions, whereas for thickness, only 17 were found to be correlated, and for thickness-derived structural graphs, 0 regions were correlated with age. Results for global mean measurements, with p-values computed with the same models, is shown in Figure 4. To recover the regression coefficient for age, with power 0.99 and alpha level 0.015, we would need 58 subjects using our method; 223 using cortical thickness; and 506 using cortical thickness-derived graph metrics.

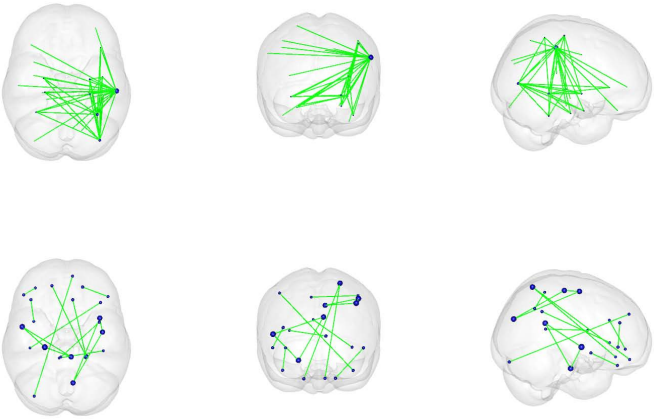


Fig. 3. Correlation between nodes (thresholded to reveal only top 2% of edges), overlaid on MNI brain. Top: Representative subject graph using our method. Bottom: Representative subject graph using cortical thickness-derived graph.

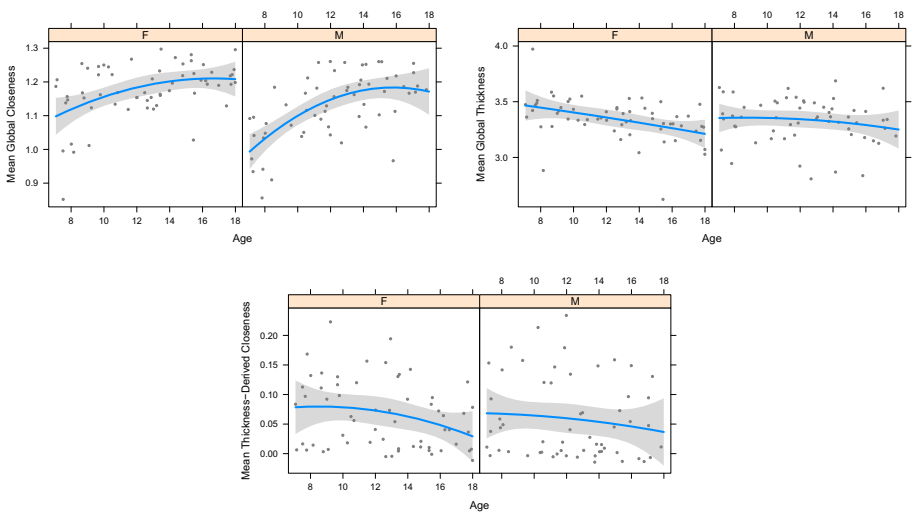


Fig. 4. Upper left: Mean closeness vs. age using our method; p-value (details in text) 5.46×10^{-8}. Upper right: Mean thickness vs. age, p-value 0.03. Bottom: Mean thickness-derived graph closeness vs. age, p-value 0.29.

4 Conclusion

We have presented a principled and closed-form method to generate single-subject cortical graphs and shown that the graphs are more sensitive to age changes than cortical thickness or cortical thickness-derived graphs are. This method has only one free parameter and shows a biologically meaningful trend with age. The method may also be used for tracking cortical changes in Alzheimer's disease and other neurodegenerative conditions.

References

1. Alexander-Bloch, A., Giedd, J.N., Bullmore, E.: Imaging structural co-variance between human brain regions. Nature Reviews Neuroscience 14(5), 322–336 (2013)
2. Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C.: A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage 54(3), 2033–2044 (2011)
3. Bassett, D.S., Bullmore, E., Verchinski, B.A., Mattay, V.S., Weinberger, D.R., Meyer-Lindenberg, A.: Hierarchical organization of human cortical networks in health and schizophrenia. The Journal of Neuroscience 28(37), 9239–9248 (2008)
4. Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal 2006. Complex Systems 1695, 1–9 (2006)
5. Dai, D., He, H., Vogelstein, J., Hou, Z.: Network-based classification using cortical thickness of AD patients. In: Suzuki, K., Wang, F., Shen, D., Yan, P. (eds.) MLMI 2011. LNCS, vol. 7009, pp. 193–200. Springer, Heidelberg (2011)
6. Greicius, M.D., Krasnow, B., Reiss, A.L., Menon, V.: Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. Proceedings of the National Academy of Sciences 100(1), 253–258 (2003)
7. Greicius, M.D., Supekar, K., Menon, V., Dougherty, R.F.: Resting-state functional connectivity reflects structural connectivity in the default mode network. Cerebral Cortex 19(1), 72–78 (2009)
8. Jafari-Khouzani, K., Soltanian-Zadeh, H.: Radon transform orientation estimation for rotation invariant texture analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(6), 1004–1008 (2005)
9. Kabsch, W.: A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A 32(5), 922–923 (1976)
10. Raj, A., Mueller, S.G., Young, K., Laxer, K.D., Weiner, M.: Network-level analysis of cortical thickness of the epileptic brain. NeuroImage 52(4), 1302–1313 (2010)
11. Segall, J.M., Allen, E.A., Jung, R.E., Erhardt, E.B., Arja, S.K., Kiehl, K., Calhoun, V.D.: Correspondence between structure and function in the human brain at rest. Frontiers in Neuroinformatics 6, 10 (2012)
12. Tijms, B.M., Seris, P., Willshaw, D.J., Lawrie, S.M.: Similarity-based extraction of individual networks from gray matter MRI scans. Cerebral Cortex 22(7), 1530–1541 (2012)
13. Varoquaux, G., Craddock, R.C.: Learning and comparing functional connectomes across subjects. NeuroImage 80, 405–415 (2013)

T_2-Relaxometry for Myelin Water Fraction Extraction Using Wald Distribution and Extended Phase Graph

Alireza Akhondi-Asl, Onur Afacan, Robert V. Mulkern, and Simon K. Warfield

Computational Radiology Laboratory, Boston Children's Hospital, and Harvard Medical School 300 Longwood Ave. Boston MA 02115 USA

Abstract. Quantitative assessment of myelin density in the white matter is an emerging tool for neurodegenerative disease related studies such as multiple sclerosis and Schizophrenia. For the last two decades, T_2 relaxometry based on multi-exponential fitting to a single slice multi-echo sequence has been the most common MRI technique for myelin water fraction (MWF) mapping, where the short T_2 is associated with myelin water. However, modeling the spectrum of the relaxations as the sum of large number of impulse functions with unknown amplitudes makes the accuracy and robustness of the estimated MWF's questionable. In this paper, we introduce a novel model with small number of parameters to simultaneously characterize transverse relaxation rate spectrum and B_1 inhomogeneity at each voxel. We use mixture of three Wald distributions with unknown mixture weights, mean and shape parameters to represent the distribution of the relative amount of water in between myelin sheets, tissue water, and cerebrospinal fluid. The parameters of the model are estimated using the variable projection method and are used to extract the MWF at each voxel. In addition, we use Extended Phase Graph (EPG) method to compensate for the stimulated echoes caused by B_1 inhomogeneity. To validate our model, synthetic and real brain experiments were conducted where we have compared our novel algorithm with the non-negative least squares (NNLS) as the state-of-the-art technique in the literature. Our results indicate that we can estimate MWF map with substantially higher accuracy as compared to the NNLS method.

Keywords: T_2 Relaxometry, MWF, EPG, Wald, Variable projection.

1 Introduction

Myelin is a layer of dielectric material derived mainly from lipids that form a sheath around neuronal axons and is well known to be crucial to support brain function [1]. Myelin-related disorders affect an estimated 3 million people around the world where this number is increasing every year. As such, the development of myelin imaging holds out the potential of providing pathologically specific quantitative information about myelin content. Among myelin imaging techniques, T_2 relaxometry is the most advantageous and effective non-invasive MRI approach

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 145–152, 2014.

for measuring alterations in myelin water content. The rationale is that the water molecules bound between myelin sheets, tissue water, and cerebrospinal fluid (CSF) have short, medium, and long T_2 relaxation times, respectively. Consequently, the fraction of water molecules with fast decay corresponds directly to the density of the myelin at each voxel. Therefore, by measuring the MR signal for multiple echo times and forming an estimate of the distribution of relaxation rates at each voxel, the fraction of water molecules characterized by fast decay can be estimated. The most well established approach for imaging of this T_2 decay is a Carr-Purcell-Meiboom-Gill (CPMG) sequence that collects many spin echo samples of the T_2 decay curve [2,3]. The standard CPMG sequence can be extended to a multi-slice CPMG sequence by changing the 180 degrees RF pulse to a slice selective RF pulses. Multi-slice T_2 CPMG acquires several slices simultaneously and allow dramatic acceleration of the acquisition [4].

Conventionally, the myelin-bound water and free water fractions are identified by fitting a discrete mixture of impulse functions, each centered at pre-specified T_2 values across the range of anticipated T_2 values. Each one of the impulse functions represents a single relaxation rate. A linear weight for each impulse function is fit to the multi-echo T_2 data via non-negative least squares (NNLS). The fraction of myelin-bound water is computed by summing the weights for all of the short components (below 50ms), and dividing by the total weight for all of the components [5,6]. Recently, an extension of the NNLS approach is introduced where Extended Phase Graph (EPG) method is used to model the imperfect refocusing in CPMG based sequences. EPG method can be used for the precise calculation of observed echoes as the function of flip-angle, T_1, T_2, and echo time [7]. The authors optimize the flip angle and weights in a two-step optimization process where in the first step they estimate the flip angle and in the second optimization stage the estimated flip angle is used to estimate the weights of the EPG functions. However, this approach of fitting a discrete mixture of impulse basis functions fails to exploit the continuity of the true distribution of T_2 in the tissue.

In this paper, we have developed an alternative representation in which we use a finite mixture of continuous distributions to describe the complete T_2 spectrum. The fraction of the myelin-bound water is the area under the fast component curve divided by the total area of each component curve. This representation has the specific advantage that the number of parameters that must be estimated from the data is much smaller. We simultaneously estimate 3 parameters per component and the flip angle, for a total of ten parameters, where the NNLS approaches estimates more than 40 parameters from 32 spin echoes.

This approach, which uses a more physically realistic model of the signal, is also easier to estimate and leads to less noisy MWF estimates. The model we have developed for the T_2 distribution of each component is the Wald distribution, which has parameters of volume of occupancy, mean and shape that completely characterize the distribution [8]. The Wald distribution has a Gaussian-like distribution with positive support which makes it suitable for the representation of transverse relaxation rate distribution. Robust and reliable estimation of the

parameters of a mixture of Wald distributions can be achieved with a well-known technique called the variable projection method, which allows us to rapidly solve this nonlinear estimation problem [9,10]. We have compared our algorithm with a well-known approach in the literature using both synthetic and real brain images and have shown the superiority of our approach.

2 Methods

2.1 Problem Definition

In the most general form, the MR signal observed at a voxel as a function of echo time is the sum of the signals from a population of spins where each one of them contributes to the observed signal as a function of R_2. Therefore, the i-th observed signal at the echo time $t_i = i \times TE$ can be expressed as:

$$S_i = S_0 \int_0^\infty f(R_2) EPG(R_2, \theta, TE, i) dR_2 \quad R_2 > 0 \tag{1}$$

where $f(R_2)$ is the probability density function (pdf) of the relaxivity rates,$R_2 = \frac{1}{T_2}$, S_0 is a constant, $EPG(R_2, \theta, TE, i)$ is the i-th stimulated echo for the spins with the relaxation rate of R_2 ,θ is the flip angle, and TE is the interecho spacing. Since the impact of the T_1 value on the stimulated echoes is negligible in CPMG based sequences, we use a fixed $T_1 = 1s$ in all of the experiments. Without loosing generality, we can assume that the density function $f(R_2)$ can be expressed as a mixture of distributions of n components:

$$f(R_2) = \sum_{j=1}^{n} a_j f_j(R_2) \quad \sum_{j=1}^{n} a_j = 1 \tag{2}$$

where $f_j(R_2)$ is the pdf of the j-th component and $a_j \leftarrow S_0 a_j$ for simplicity. It is known that the spectrum of relaxivity rate has $n \leq 3$ Gaussian-like components [3]. There are variety of pdf's with positive support which can be used to model the distribution of the components such as truncated Gaussian, log-normal, and Wald distributions. Since the pdf should satisfy $f(R_2 = 0) = 0$, truncated Gaussian is not appropriate distributions in the general case. Here, we use Wald distribution to model $f_j(R_2)$:

$$f_j(R_2) = \left(\frac{\lambda_j}{2\pi R_2^3}\right)^{\frac{1}{2}} \exp\left(\frac{-\lambda_j}{2\mu_j^2 R_2}(R_2 - \mu_j)^2\right) \quad R_2 > 0 \tag{3}$$

where $\mu_j > 0$ and $\lambda_j > 0$ are mean and shape parameter of the distribution, respectively and $\frac{\mu_j^3}{\lambda_j}$ is the variance of the distribution. The Wald distribution has several properties similar to the normal distribution. In addition, for small standard deviations, it becomes very similar to the Gaussian distribution.

2.2 Optimization

We are interested to estimate the parameters of the Wald distributions, their mixture weights, and flip angle using the observed signals at different echo times. However, in practice, we observe y_i, a noisy version of the signal S_i. We assume that zero mean, additive white Gaussian noise is added to the signal S_i.

Let, $\{\boldsymbol{\Phi}(\boldsymbol{\alpha})\}_{i,j} = \phi_j(\mu_j, \lambda_j, \theta; t_i)$ be a matrix of size $m \times n$ where

$$\phi_j(\mu_j, \lambda_j, \theta; t_i) = \int_0^\infty \left(\frac{\lambda_j}{2\pi R_2^3}\right)^{\frac{1}{2}} \exp\left(\frac{-\lambda_j}{2\mu_j^2 R_2}(R_2 - \mu_j)^2\right) EPG(R_2, \theta, TE, i) dR_2 \tag{4}$$

and $\boldsymbol{\alpha} = (\mu_1, \lambda_1, \ldots, \mu_n, \lambda_n, \theta) \in \mathcal{R}^{2n+1}$ be the vector of the mean and shape parameters of n Wald distributions, and the flip angle. Given data $(t_i, y_i), i = 1, \ldots, m \geq 3n + 1$, we want to find set of parameters $\hat{\mathbf{a}} = (\hat{a}_1, \ldots, \hat{a}_n)$, $\hat{\boldsymbol{\alpha}} = (\hat{\mu}_1, \hat{\lambda}_1 \ldots, \hat{\mu}_k, \hat{\lambda}_k, \hat{\theta})$ which minimize the following functional:

$$r(\mathbf{a}, \boldsymbol{\alpha}) = ||\mathbf{y} - \boldsymbol{\Phi}(\boldsymbol{\alpha})\mathbf{a}||^2 = \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} a_j \phi_j(\mu_j, \lambda_j, \theta; t_i)\right)^2 \tag{5}$$

where $\mathbf{a} = (a_1, \ldots, a_n) \in \mathcal{R}^n$ and $\mathbf{y} = (y_1, \ldots, y_m) \in \mathcal{R}^m$ are the vectors of mixture weights and noisy observations, respectively.

This functional can be optimized using any Non-linear least squares (NLLS) optimization algorithm. However, since, it has separable NLLS formulation, it is possible to use a smart approach to improve the performance of the optimization [10]. Let us assume that the nonlinear parameters $\boldsymbol{\alpha}$ are known. Therefore, the linear parameters which satisfies the minimal least square solution can be written as $\hat{\mathbf{a}} = \boldsymbol{\Phi}(\boldsymbol{\alpha})^+ \mathbf{y}$ where the matrix $\boldsymbol{\Phi}(\boldsymbol{\alpha})^+$ is the Moore-Penrose generalized inverse of $\boldsymbol{\Phi}(\boldsymbol{\alpha})$. By replacing $\hat{\mathbf{a}} = \boldsymbol{\Phi}(\boldsymbol{\alpha})^+ \mathbf{y}$, the variable projection functional can be constructed:

$$\min_{\mathbf{a}} r(\mathbf{a}, \boldsymbol{\alpha}) = r(\hat{\mathbf{a}}, \boldsymbol{\alpha}) = ||\left(\mathcal{I} - \boldsymbol{\Phi}(\boldsymbol{\alpha})\boldsymbol{\Phi}^+(\boldsymbol{\alpha})\right)\mathbf{y}||^2 = ||\mathcal{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\alpha})}\mathbf{y}||^2 = r_2(\boldsymbol{\alpha}) \tag{6}$$

where the matrix $\mathcal{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\alpha})} = \mathcal{I} - \boldsymbol{\Phi}(\boldsymbol{\alpha})\boldsymbol{\Phi}^+(\boldsymbol{\alpha})$ is the projector on the orthogonal complement of the column space of $\boldsymbol{\Phi}(\boldsymbol{\alpha})$.

This indicates that we can first optimize nonlinear parameters by eliminating the linear parameters from the optimization problem. Then, we can use the obtained non-linear parameters to estimate the linear ones using the minimal least square solution. This approach has shown to be very effective in cases where the number of linear parameters is substantial. To optimize the variable projection functional without analytic derivatives, one needs to iteratively compute $r_2(\boldsymbol{\alpha})$. However, to improve the performance of the algorithm, it is possible to use the analytical derivatives in a Levenberg-Marquardt type NNLS solver. Let $\mathbf{D}_k$ be a matrix of size $m \times n$ where $\{\mathbf{D}_k\}_{i,j} = \frac{\partial \phi_j(\mu_j, \lambda_j; t_i)}{\partial \alpha_k}$ $k = 1 \ldots 2n + 1$. It is

known that $\mathbf{J}_k$, k-th column of $m \times 2n+1$ Jacobian matrix $\mathbf{J} = \frac{\partial \mathcal{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\alpha})}\mathbf{y}}{\partial \alpha_k}$, can be computed as [10,9]:

$$\mathbf{J}_k = -\mathcal{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\alpha})}\mathbf{D}_k\boldsymbol{\Phi}^{+}\mathbf{y} - (\boldsymbol{\Phi}^{+})^T\mathbf{D}_k^T\mathcal{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\alpha})}\mathbf{y} \tag{7}$$

Therefore, to derive Jacobian matrix, we only need to provide $\frac{\partial \phi_j(\mu_j,\lambda_j,\theta;t_i)}{\partial \alpha_k}$ which can be computed using the following relation:

$$\frac{\partial \phi_j(\mu_j,\lambda_j,\theta;t_i)}{\partial \alpha_k} = \tag{8}$$

$$\begin{cases} \int_0^{\infty} \left(\frac{\lambda_j}{2\pi R_2^3}\right)^{\frac{1}{2}} \exp\left(\frac{-\lambda_j(R_2-\mu_j)^2}{2\mu_j^2 R_2}\right) \frac{\lambda_j R_2(R_2-\mu_j)}{\mu_j^3 R_2} EPG(R_2,\theta,TE,i)dR_2 & k = 2j-1 \\ \int_0^{\infty} \exp\left(\frac{-\lambda_j(R_2-\mu_j)^2}{2\mu_j^2 R_2}\right) \frac{1-\frac{\lambda_j(R_2-\mu_j)^2}{\mu_j^2 R_2}}{2(\lambda_j 2\pi R_2^3)^{\frac{1}{2}}} EPG(R_2,\theta,TE,i)dR_2 & k = 2j \\ \int_0^{\infty} \left(\frac{\lambda_j}{2\pi R_2^3}\right)^{\frac{1}{2}} \exp\left(\frac{-\lambda_j}{2\mu_j^2 R_2}(R_2-\mu_j)^2\right) \frac{\partial EPG(R_2,\theta,TE,i)}{\partial \theta} dR_2 & k = 2J+j \\ 0 & otherwise \end{cases}$$

where $\frac{\partial EPG(R_2,\theta,TE,i)}{\partial \theta}$ can be computed recursively.

3 Results

3.1 Synthetic Data

We use synthetic data with the known ground truth to evaluate our developed method and the NNLS algorithm. To demonstrate the impact of modeling R_2 spectrum with a mixture of continuous distribution, we also evaluate a modified version of our model where the Wald distribution is replaced with the impulse function. Mixture of three Wald distributions is considered as the ground truth with the peaks at 50Hz, 10Hz, and 1Hz, the shape parameters of $600Hz$, $400Hz$, and $300Hz$, and weights of 0.2,0.6, and 0.1, respectively.

Figure 1.a shows the normalized Cramer-Rao lower bound (CRLB) of MWF estimated using our method where we normalized the CRLB by the true MWF value. We have computed CRLB for all combinations of 5 equally spaced SNRs between $30dB$ and $50dB$ and 5 equally spaced flip angles between 120^o and 200^o. This figure indicates that MWF can be estimated with very high accuracy in a clinical imaging scenario (SNR=$40dB$ and flip angle larger than 120^o). Next, we evaluate our optimization algorithm for the same range of SNRs and flip angles. We initialize our method with three Wald distributions with means at $30ms$, $90ms$, and $1500ms$ and with shape parameters of $500Hz$. To have a more robust optimization, we use constraints on the mean and shape parameters of our model. We assume that the mean of the components are in the range of $15-40ms$, $60-120ms$, $200-2000ms$ and the shape parameters are between $0.01-10kHz$. These are reasonable numbers without any strong assumption about the R_2 spectrum. We observe 32 echoes at the range of $8-256ms$ and we repeat each experiment 1000 times at each SNR. Figure 1.b shows the relative

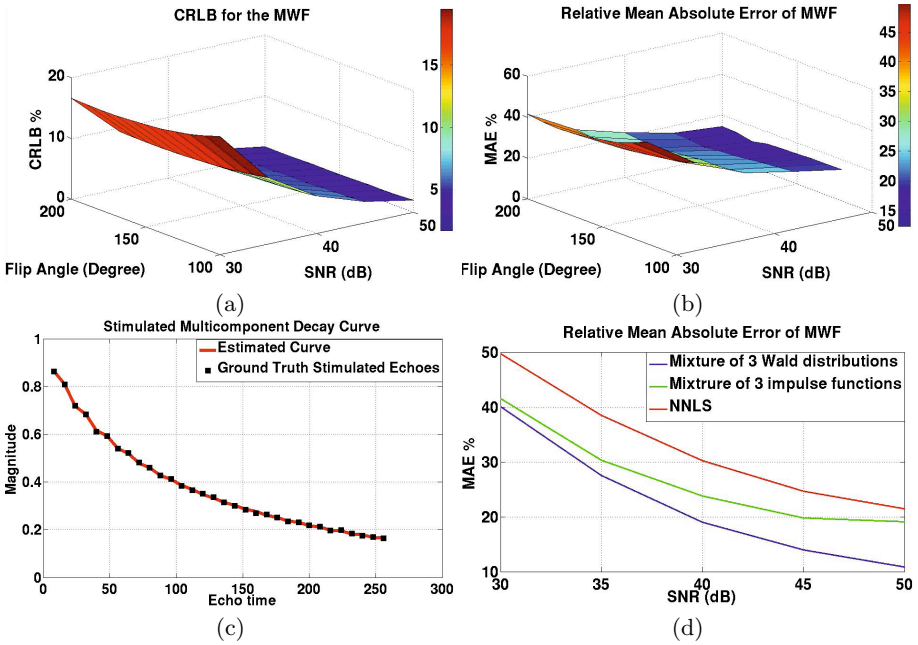
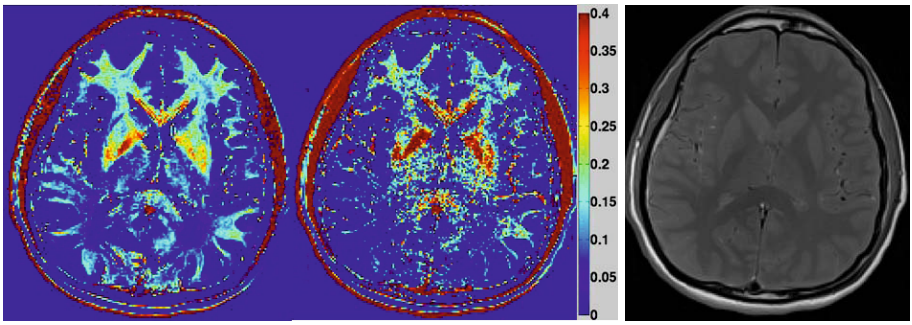


Fig. 1. Quantitative evaluation of mixture of Wald distributions. (a) CRLB of MWF estimation using mixture of Wald distributions where the standard deviation is normalized by the true MWF value. (b) Relative MAE of Wald distribution for a range of SNRs and flip angles. (c) Estimated stimulated echoes using the mixture of Wald distributions for $SNR = 45dB$ and flip angle of 150^o. (d) Relative MAE of the three methods for the range of SNRs and flip angle of 180^o. The results indicate that we can estimate the MWF accurately for the practical flip angles and SNRs. It can also be seen that our method has lower MAE as compared to the other methods for all the SNRs in the range of $30 - 50dB$.

mean absolute error (MAE) of our method at different SNRs and flip angles. As seen, for the practical flip angles and SNRs, our method estimates the MWF with very small error. Figure 1.c shows a stimulated multicomponent decay curve for $SNR = 45dB$ and flip angle of 150^o and the estimated curve using our algorithm. This figure shows that we can accurately estimate both the R_2 spectrum parameters and the flip angle, as the stimulated echoes are estimated with very high accuracy. Finally, we compare the performance of our method, NNLS, and our modified model with three impulse functions for a range of SNRs and flip angle of 180^o. For this experiment, we use inverse gamma distribution to generate the ground truth with the same mean, standard deviation, and fractions of the previous experiments. In this way, we will have a fair comparison between the methods, as a different distribution is utilized to generate the ground truth spectrum. For NNLS method we estimate the amplitude of 50 impulse functions logarithmically spaced within the range of $15ms$ and $2s$. For our model with three impulse functions, we use the initialization and constraints of the mean

parameters of the Wald distributions. Figure 1.d shows the relative MAE of the estimation of MWF using three different methods. It can be seen that our method has substantially smaller error as compared to the other methods for the whole SNR range. This indicates that our approach can produce the performance of NNLS method in the substantially lower SNR.



(a) (b)

Fig. 2. Qualitative comparison of algorithms. Estimated MWF map using mixture of Wald distribution (a-left) and NNLS (a-right) indicates that the estimated map using our approach is less noisy as compared to the NNLS algorithm. (b) The image of the first echo.

3.2 Real Brain MRI

In addition, our method and the NNLS algorithm were tested on 10 volunteers. T_2 relaxation measurements were performed on a 3T Siemens TRIO scanner with a multi-echo CPMG sequence acquiring 32 echoes with an echo spacing of 9 ms. A 21cm FOV was used with a matrix size of 192x192 (in plane resolution of 1.1mm) and the total scan time was 9 minutes and 41 seconds for acquisition of 3mm thick slices. Parameter initialization for each of the estimation procedures was the same as the simulation experiment. For the NNLS algorithm the two-step optimization approach in [6] was used to correct for the stimulated echoes. Figure 2.a shows the MWF mapping of one slice of a subject using our method and NNLS. For both models the components with T_2 shorter than 50ms are used to estimate the MWF. The results show that our approach estimated the MWF more accurately as compared to the NNLS method, since the MWF map is smoother and less noisy. Figure 2.b shows the image of the first echo.

4 Conclusions

In this paper, we have introduced a novel model to represent the spectrum of the relaxation rate at each voxel. To this end, we have utilized a mixture of three Wald distributions with unknown mixture weights, means and shape parameters. We also used EPG method to model stimulated echoes. Finally, we

have utilized variable projection method to optimize the unknown parameters and used the estimated mean and mixture weights to identify the MWF at each voxel. We have used both synthetic and real brain images for the validation of our method. In addition, we have compared our method with the state-of-the-art MWF extraction algorithm and showed the superiority of our method.

Acknowledgment. This research was supported in part by NIH grants R01 EB013248, R01 LM010033, R01 NS079788, R42 MH086984, P30 HD018655, R03 EB008680 and by a research grant from Boston Children's Hospital Translational Research Program.

References

1. Nave, K.A.: Myelination and support of axonal integrity by glia. Nature 468(7321), 244–252 (2010)
2. Kolind, S.H., Madler, B., Fischer, S., Li, D.K., MacKay, A.L.: Myelin water imaging: implementation and development at 3.0 T and comparison to 1.5 T measurements. Magnetic Resonance in Medicine 62(1), 106–115 (2009)
3. Whittall, K.P., MacKay, A.L.: Quantitative interpretation of NMR relaxation data. Journal of Magnetic Resonance 84(1), 134–152 (1969, 1989)
4. Oh, J., Han, E.T., Lee, M.C., Nelson, S.J., Pelletier, D.: Multislice brain myelin water fractions at 3T in multiple sclerosis. Journal of Neuroimaging 17(2), 156–163 (2007)
5. Mackay, A., Whittall, K., Adler, J., Li, D., Paty, D., Graeb, D.: In vivo visualization of myelin water in brain by magnetic resonance. Magnetic Resonance in Medicine 31(6), 673–677 (1994)
6. Prasloski, T., Rauscher, A., MacKay, A.L., Hodgson, M., Vavasour, I.M., Laule, C., Mädler, B.: Rapid whole cerebrum myelin water imaging using a 3D GRASE sequence. Neuroimage (2012)
7. Hennig, J.: Echoes-how to generate, recognize, use or avoid them in MR-imaging sequences. part I: Fundamental and not so fundamental properties of spin echoes. Concepts in Magnetic Resonance 3(3), 125–143 (1991)
8. Seshadri, V.: The inverse Gaussian distribution: a case study in exponential families. Clarendon Press Oxford (1993)
9. O'Leary, D.P., Rust, B.W.: Variable projection for nonlinear least squares problems. Computational Optimization and Applications, 1–15 (2012)
10. Golub, G., Pereyra, V.: Separable nonlinear least squares: the variable projection method and its applications. Inverse Problems 19(2), R1 (2003)

Compact and Informative Representation of Functional Connectivity for Predictive Modeling

Raif M. Rustamov[1], David Romano[2], Allan L. Reiss[2], and Leonidas J. Guibas[1]

[1] Computer Science Department, Stanford University, Stanford, CA
[2] Center for Interdisciplinary Brain Sciences Research, Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA

Abstract. Resting state functional connectivity holds great potential for diagnostic prediction of neurological and psychiatric illness. This paper introduces a compact and information-rich representation of connectivity that is geared directly towards predictive modeling. Our representation does not require a priori identification of localized regions of interest, yet provides a mechanism for interpretation of classifier weights. Experiments confirm increased accuracy associated with our representation and yield interpretations consistent with known physiology.

1 Introduction

Resting state functional magnetic resonance imaging (fMRI), in conjunction with multivariate pattern analyses, holds great promise for diagnostic prediction of neurological and psychiatric illness [2]. For accurate predictive modeling, it is necessary to have compact representations of functional connectivity. Such representations are usually obtained by a judicious choice of nodes for assembling the correlation matrix/connectivity network. Parcellation schemes based on anatomical and/or functional features are used to yield regions of interest (ROIs) that are identified as nodes of the network [10]. For network-based analyses of anatomical and functional connectivity it is crucial to use nodes derived from spatially localized ROIs that are functionally and biologically meaningful [5]. On the other hand, for predictive modeling it is desirable to maximize the information content of the reduced representation. For example, a priori identification of ROIs may obscure more subtle and complex phenomena that cross the boundaries of ROIs and thus may lead to suboptimal prediction accuracy.

The goal of our work is to introduce a compact and informative representation of functional connectivity that is geared directly towards predictive modeling and does not require a priori identification of localized ROIs. Our approach is based on the observation that a localized ROI can be fully captured by its indicator function, which is simply a particular type of spatial map (i.e., real-valued functions on the collection of voxels). The set of all possible spatial maps is a very high-dimensional vector space, but since fMRI data is already subject to smoothing, both in its acquisition and preprocessing, we choose to restrict to the subspace of spatial maps exhibiting some smoothness. This subspace can

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 153–160, 2014.

be approximated by the span of low-frequency eigenvectors of an appropriate graph Laplacian (in the same way as Fourier basis provides building blocks for time signals, the Laplacian eigenvectors provide a basis for spatial signals). Therefore, we build our representation from the spatial maps associated to low-frequency eigenvectors, which replace ROIs as nodes of network. Namely, these distributed and overlapping spatial maps are combined with blood oxygenation level dependent (BOLD) signal to obtain the corresponding representative time series. The matrix of correlations between these time series are then computed, giving the sought representation of connectivity.

The proposed approach has a number of advantages. First, it leads to a compact representation of functional connectivity that is hierarchical. Indeed, the Laplacian eigenvectors are naturally ordered by their smoothness and a number of smoother eigenvectors can be retained to obtain a connectivity matrix of desired size. Second, the proposed representation is informative in that it allows approximate reconstruction of correlations between any pair of traditional ROIs, even if these ROIs were not specified a priori. This explains why our approach is effective: in a sense, traditional ROIs are subsumed by our approach and a classifier applied to our representation is able to learn any information that could be extracted from traditional ROIs. Finally, our representation is interpretable—it provides a simple mechanism for mapping the weights learned by linear classifiers back to the brain, allowing a detailed understanding of the predictive model.

2 Methods

Let X be the $n_{\mathrm{TR}} \times n_{\mathrm{vox}}$ matrix of the voxel-wise z-scored corrected BOLD signal acquired in n_{vox} grey matter voxels over n_{TR} time points. The matrix $C = X^\top X$ captures the correlations between all pairs of voxels, but is impractically large. Our goal is to obtain a computationally tractable reduced representation of C.

Motivation. To motivate our representation, let us analyze the commonly used ROI-based approaches. For i-th ROI, consider an $n_{\mathrm{vox}} \times 1$ column vector whose entries correspond to voxels, with an entry of 1 for voxels that belongs to the ROI, and 0 otherwise. Let us scale this vector so that its entries sum to 1, and denote this normalized vector by ϕ^i; this is the indicator vector of the ROI. Note that the product $X\phi^i$ gives the averaged BOLD time series over the i-th ROI, and so the product $(X\phi^i)^\top X\phi^j = \phi^{i\top} X^\top X\phi^j = \phi^{i\top} C\phi^j$ gives the un-normalized correlation between i-th and j-th ROIs. Denote by Φ the $n_{\mathrm{vox}} \times n_{\mathrm{ROI}}$ matrix whose i-th column is the vector ϕ^i. Now the matrix $\Phi^\top C\Phi$ can be seen to be an $n_{\mathrm{ROI}} \times n_{\mathrm{ROI}}$ matrix of un-normalized correlations between all the ROIs.

This process can be interpreted in the following manner. The large matrix C defines a dot product (namely, $\langle f, g\rangle = f^\top Cg$) over the space of all spatial maps. The goal is to capture information about this dot product, and so about connectivity, in a compact manner. The ROI based approach achieves this by restricting the dot product to a much smaller subspace of spatial maps. Namely, the subspace in question is $\mathrm{span}\{\phi^1, \phi^2, ..., \phi^{n_{\mathrm{ROI}}}\}$, and the matrix of the restricted dot product is precisely the matrix $\Phi^\top C\Phi$ above.

Proposed Approach. In the light of the discussion above, the gist of our approach is to choose a different subspace of spatial maps for restricting the dot product. For multivariate pattern analyses it is desirable to maximize the information content of the reduced representation and this can be achieved by choosing a subspace that offers strong approximation properties. Since spatial smoothing is applied to BOLD signal during preprocessing, it is natural to restrict the dot product to a subspace of smooth spatial maps.

We propose to use the subspace spanned by the low spatial frequency eigenvectors of an appropriate graph Laplacian. Here, we consider the grey matter template voxels as nodes of a graph, and introduce an edge between voxels that share a face. By solving the eigenvalue problem $L\psi^i = \lambda_i \psi^i$, where L is the $n_{\text{vox}} \times n_{\text{vox}}$ graph Laplacian [4], we obtain an orthonormal basis $\{\psi^i\}_{i=1}^{n_{\text{vox}}}$. Assuming the eigenvectors are ordered by increasing eigenvalue, it can be shown (c.f. [1] Sec. 5.2) that in some precise mathematical sense the eigenvector ψ^1 is the smoothest spatial map, ψ^2 the next smoothest, and so on. Since we would like to obtain low-dimensional subspace of smooth spatial maps, we consider the subspace spanned by the first n_{ev} eigenvectors. Denoting by Ψ an $n_{\text{vox}} \times n_{\text{ev}}$ matrix whose i-th column is ψ^i, our reduced representation of the voxel-wise connectivity matrix C is given by $D = \Psi^\top C \Psi$ which is a matrix of size $n_{\text{ev}} \times n_{\text{ev}}$. In practice, we first compute $X\psi^i$ for $i = 1, 2, ..., n_{\text{ev}}$, which gives the BOLD time series weighted by the spatial map ψ^i, and then we set the entries $D_{ij} = (X\psi^i)^\top X\psi^j$.

Informativeness. The proposed representation is informative in the sense that it allows approximate reconstruction of correlations between two given ROIs. Indeed, let ϕ^1 and ϕ^2 be the indicator vectors of the two ROIs. As discussed in *Motivation*, the un-normalized correlation between the ROIs is given by $\phi^{1\top} C \phi^2$. Let us write the vectors ϕ^1 and ϕ^2 in terms of the eigenvector basis: $\phi^1 = \alpha_1\psi^1 + \alpha_2\psi^2 + \cdots = \Psi\alpha$ and $\phi^2 = \beta_1\psi^1 + \beta_2\psi^2 + \cdots = \Psi\beta$; here α (resp. β) is a column-vector with entries α_i (resp. β_i). Now we have, $\phi^{1\top} C \phi^2 = (\Psi\alpha)^\top C(\Psi\beta) = \alpha^\top \Psi^\top C \Psi \beta = \alpha^\top D \beta$, where we used $D = \Psi^\top C \Psi$. We can compute the correlations between the ROIs by appropriate normalization, namely:

$$\text{corr}(ROI_1, ROI_2) = \frac{\alpha^\top D \beta}{(\alpha^\top D \alpha)^{1/2} (\beta^\top D \beta)^{1/2}} \tag{1}$$

Note that since we have actually truncated the eigenvector basis, this equality holds only approximately. This approximate reconstruction property shows that despite its compactness, our representation is able to capture connectivity in a richly informative manner, which is crucial for accurate predictive modeling.

Interpretability. The proposed representation is interpretable in the sense that the feature weightings obtained from linear classification algorithms can be mapped back to locations in the brain. Let $\hat{y} = \text{sign}(b + \sum_{i,j} W_{ij} D_{ij})$ be the prediction model where W_{ij} are the weights and b is the bias; this form can capture commonly used linear classifiers such as Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM). Using $D = \Psi^\top C \Psi$, we can rewrite $\sum_{i,j} W_{ij} D_{ij} = \sum W_{ij} \sum_{p,q} \Psi_{ip} C_{pq} \Psi_{jq} = \sum_{p,q} C_{pq} \sum_{i,j} W_{ij} \Psi_{ip} \Psi_{jq}$; here i, j run over eigenvector indices, and p, q run over voxels. Thus, we can say that the

coefficient of C_{pq} given by $R_{pq} = \sum_{i,j} W_{ij}\Psi_{ip}\Psi_{jq}$ captures the contribution of connectivity between voxels p and q. It is easy to see that $R = \Psi W \Psi^\top$.

The matrix R is impractically big, but from it, we can derive a quantity that measures the importance of the connectivity of a given voxel for classification. For a voxel p, consider quantity $\omega_p = \sum_q R_{pq}^2$; note that squaring the coefficients is appropriate here, as we would like to amplify importance of voxels whose connectivity coefficients are larger. Letting ω be the vector with entries ω_p, we have:

$$\omega = \mathrm{diag}(R^\top R) = \mathrm{diag}(\Psi W \Psi^\top \Psi W \Psi^\top) = \mathrm{diag}(\Psi W W \Psi^\top)\,, \tag{2}$$

where the last expression is nothing but the vector of row sums of squared entries of $W\Psi^\top$. Here we used that $\Psi^\top\Psi$ is the identity matrix since the full eigenvector basis is orthonormal; in practice since we truncate the basis, the equality above is only an approximation, but it still gives a measure of how much each voxel's connectivity contributes to the classification.

Further Interpretability. Deeper insight into the learned prediction model can be obtained if one uses so-called bilinear classifiers. To motivate their use, recall that a quantity of the form $u^\top D v$ captures the un-normalized correlation between BOLD signals weighted by spatial maps Ψu and Ψv (c.f. *Informativeness*). Assuming that a limited number of such pair-wise correlations should suffice for prediction, one can seek a classifier of the form $\hat{y} = \mathrm{sign}(\sum_k \sigma_k u^{k\top} D v^k)$. Here, we let k index the pair-wise terms; without loss of generality, we can assume that vectors u^k and v^k have unit lengths, and to compensate, we introduce a scalar factor σ_k. With this notation, one can think of σ_k as the importance of each pair-wise connectivity $u^{k\top} D v^k$ to classification task. It is easy to show that we can rewrite $\hat{y} = \mathrm{sign}(\sum_k \sigma_k u^{k\top} D v^k) = \mathrm{sign}(\sum_{i,j} W_{ij} D_{ij})$. Here $W = \sum_k \sigma_k u^{k\top} D v^k$, which is exactly the singular value decomposition (SVD) of W, and now it becomes clear that our assumption is equivalent to W being a low-rank matrix. Classifiers satisfying low-rank constraints have been studied in the machine learning literature, and we will use Bilinear SVM [8] in our experiments. In practice, we find that the learned W is symmetric up to a small error (presumably since D is symmetric), so we apply SVD to the symmetrized matrix $W + W^\top$. Due to symmetry, we get $u^k = v^k$, and for visualization we depict the spatial maps Ψu^k ordered by the decreasing singular value σ_k.

3 Experiments and Results

Data and Preprocessing. We utilize the public dataset[1] from Beijing Eyes Open/Closed study [6]. The study included 48 healthy subjects who underwent three resting state scanning sessions. In the first session, all of the subjects were scanned with their eyes closed; the remaining two sessions were scanned one with eyes open (EO) and the other with eyes closed (EC) in a random order counterbalanced across subjects. Since our goal is to conduct analyses for distinguishing EO and EC, we use the data from the second and third sessions only.

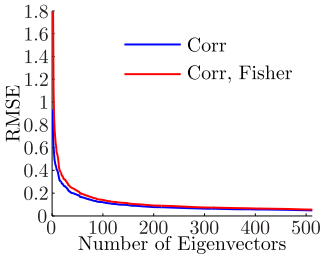
[1] `http://fcon_1000.projects.nitrc.org/indi/retro/BeijingEOEC.html`

The data is processed in the following manner. As in [6], we start by discarding the first 10 volumes of each session. Next, we use SPM8 (The Wellcome Trust Centre for Neuroimaging) for spatial preprocessing. We then use the CONN Toolbox [11] with default settings to obtain corrected BOLD signal. The corrected BOLD signal is z-scored at every voxel, giving the final normalized BOLD signal. This pipeline failed to process two of the subjects, and thus all of our experiments are based on the remaining 46 subjects.

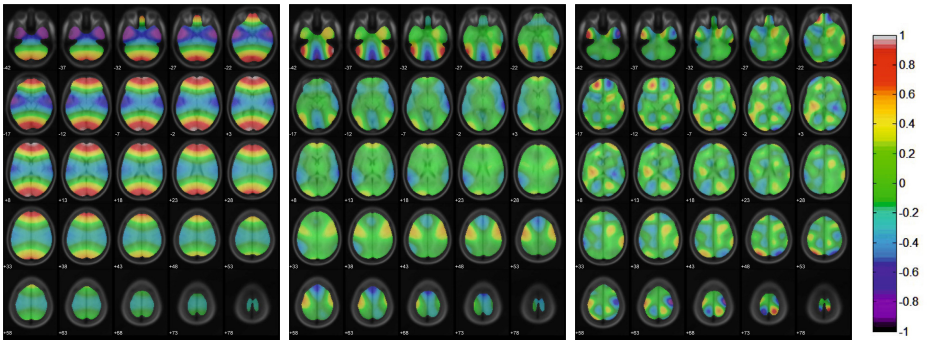
To obtain the proposed representation of resting state connectivity, we first compute a number of low-frequency eigenvectors of the grey matter voxel grid Laplacian. Next, we normalize the eigenvectors to have unit ℓ_1-norm. The normalized BOLD signal is then weighted by the eigenvectors, and the resulting time series are used to compute the dot product and correlation matrices. For comparison, we also compute connectivity matrices based on the commonly used Automated Anatomical Labeling (AAL) atlas [9] with 90 ROIs.

Experimental Results. Fig. 1 depicts the eigenvectors of the grey matter template graph Laplacian that correspond to the indices $5, 25$ and 75. For ease of visualization, the eigenvectors are normalized to have maximum absolute value of 1. The corresponding spatial maps can be seen to have varying weights, both positive and negative, and are distributed across the entire grey matter volume. We stress that no physiological meaning is associated with these eigenvectors.

The inset plot on the right depicts the error incurred when reconstructing correlations between AAL ROIs from our representation. In this experiment, we compute all unique non-diagonal correlations between AAL ROIs for one subject. Next, these ground truth values are approximated using Eq. (1) from our representation. We varied the number of eigenvectors used in our representation from 1 to 512. The error is measured by the root-mean-square error (RMSE) over pairs of ROIs. The plot shows the RMSE both for correlations and their Fisher z-transformed versions. As expected, the reconstruction error decreases with the growing number of eigenvectors.

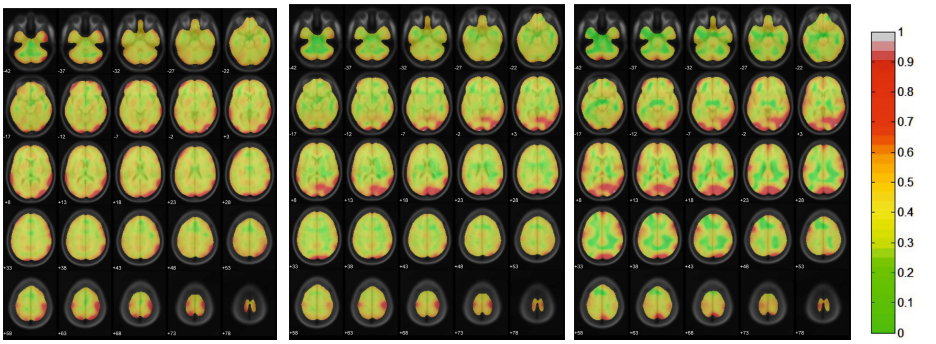


Next, we exemplify our representation in a multivariate analysis task, where our goal is to train classifiers that use resting state connectivity for predicting whether the subject had their eyes open or closed for a particular scan. We train three different classifiers: linear discriminant analysis (LDA) as implemented in MATLAB, with default settings; a linear support vector machine (SVM) as implemented in libsvm [3] with default parameters; and Bilinear SVM (maximum rank set to $d = 8$) with our in-house implementation that directly follows the algorithm description in [8]. LDA and SVM input a vector containing all of the unique entries in connectivity matrices; Bilinear SVM inputs the connectivity matrices directly. Each feature input to the linear SVM is z-scored over all subjects as this improves the performance; LDA is completely insensitive to z-scoring; the inputs to Bilinear SVM are not z-scored as this would potentially destroy the low rank structure of the coefficient matrix.



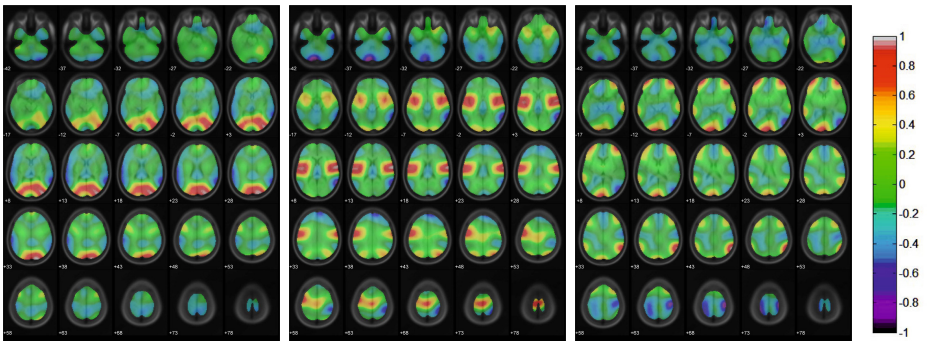
(a) (b) (c)

Fig. 1. Brain images corresponding to the eigenvectors of index (a) 5, (b) 25, and (c) 75 of the grey matter template graph Laplacian. Eigenvectors with higher indices have higher frequencies; i.e., oscillate more frequently in the space.



(a) (b) (c)

Fig. 2. Images depicting importance of voxel connectivity (i.e., "importance" maps), computed from (a) LDA, (b) SVM, and (c) Bilinear SVM classifier weights, respectively, using Eq. (2). Please refer to the text for details.



(a) (b) (c)

Fig. 3. Images corresponding to the largest three singular values of the Bilinear SVM weights, appearing left to right. Please refer to the text for details.

Table 1. Average accuracy (%) of leave-one-out cross-validation over the two sessions.

Method	LDA		SVM		Bilinear SVM	
	corr	dot	corr	dot	corr	dot
AAL	78.26	78.26	80.44	76.09	76.08	79.35
AALrz	80.44	73.92	79.35	72.83	77.18	69.57
Proposed	90.22	89.13	91.30	86.96	85.87	86.96

The classification experiments are conducted as follows. To avoid confounding, we train and test the classifiers on each session (i.e., the second and third sessions) separately. Due to the small sample size and standard practice in the neuroimaging literature, classification performance is measured via leave-one-out cross-validation accuracy. We report average accuracy over the two sessions.

Table 1 shows the performance of these classifiers in detecting EO vs. EC for various representations of functional connectivity. We compare against two versions of AAL ROI based approach—one using the voxel-wise z-scored corrected BOLD signal (AAL) and the other using the corrected BOLD signal that is z-scored region-wise (AALrz). Since there are 90 AAL ROIs, we build our representation using 90 eigenvectors; thus, all connectivity representations in this experiment have the same size (90×90). Columns marked with "dot" use as features the un-normalized correlations (e.g. entries D_{ij} directly) and columns marked with "corr" use the usual correlation (e.g. the normalized entries $D_{ij}/\sqrt{D_{ii}D_{jj}}$). For all combinations, our proposed representation is seen to improve classification accuracy over the AAL by a non-negligible amount.

Fig. 2 depicts the overall importance maps from Eq. (2), computed using the weights from classifiers trained on the entire dataset. For visualization purposes, the maps are scaled to have maximum absolute value of 1. The regions highlighted in red correspond to the voxels whose connectivity is most helpful for classification. For example, the highlighted posterior regions coincide roughly with the visual cortex, where visual sensory data is processed. Fig. 3 depicts the spatial maps corresponding to the first three singular values of the Bilinear SVM's symmetrized weight matrix, and provides an illustration of the information content that can be extracted by our methods. The image corresponding to the first singular value, which represents the component of largest influence, shows prominently (in red) the contribution of the visual cortex to the task of discriminating between EO and EC. Furthermore, the motor cortex and orbitofrontal cortex—regions which are implicated in voluntary movement and the conscious control of attention (often referred to as executive function), respectively—are prominent (red) in the images corresponding to the second and third singular values, respectively, consistent with their secondary roles in the eyes open condition. These results are anatomically consistent with existing analyses that interrogate network differences between these two states [7,12]. Note that classifier weights should be interpreted with caution, as they can reflect properties of the methods used rather than the data they are applied to; yet the fact that they are consistent with known physiology is encouraging.

4 Conclusion

We have introduced a novel information-rich compact representation of functional connectivity based on using low-frequency Laplacian eigenvectors as the spatial maps. The resulting maps only depend on the grey matter template and so enjoy the kind of data-independent universality usually associated with atlas-based methods. Experiments confirm increased classifier accuracy with our representation and lead to plausible interpretations.

Acknowledgments. This work was supported by NIH grant T32 MH019908, ONR MURI grant N00014-13-1-0341, AFOSR grant FA9550-12-1-0372, NSF grant DMS 1228304, and the Max Planck Center for Visual Computing and Communications. Support for the publicly available Beijing EO/EC data was provided by a grant from the National Natural Science Foundation of China: 30770594 and a grant from the National High Technology Program of China (863): 2008AA02Z405.

References

1. Belkin, M., Niyogi, P.: Semi-supervised learning on riemannian manifolds. Machine Learning 56(1-3), 209–239 (2004)
2. Castellanos, F.X., Martino, A.D., Craddock, R.C., Mehta, A.D., Milham, M.P.: Clinical applications of the functional connectome. NeuroImage 80, 527–540 (2013)
3. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1 – 27:27 (2011)
4. Chung, F.R.K.: Spectral Graph Theory. American Mathematical Society (1997)
5. Fornito, A., Zalesky, A., Breakspear, M.: Graph analysis of the human connectome: Promise, progress, and pitfalls. NeuroImage 80, 426–444 (2013)
6. Liu, D., Dong, Z., Zuo, X., Wang, J., Zang, Y.: Eyes-open/eyes-closed dataset sharing for reproducibility evaluation of resting state fMRI data analysis methods. Neuroinformatics 11(4), 469–476 (2013)
7. Patriat, R., Molloy, E.K., Meier, T.B., Kirk, G.R., Nair, V.A., Meyerand, M.E., Prabhakaran, V., Birn, R.M.: The effect of resting condition on resting-state fMRI reliability and consistency: A comparison between resting with eyes open, closed, and fixated. NeuroImage 78, 463–473 (2013)
8. Pirsiavash, H., Ramanan, D., Fowlkes, C.: Bilinear classifiers for visual recognition. In: NIPS, pp. 1482–1490 (2009)
9. Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M.: Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 15(1), 273–289 (2002)
10. Varoquaux, G., Craddock, R.C.: Learning and comparing functional connectomes across subjects. NeuroImage 80, 405–415 (2013)
11. Whitfield-Gabrieli, S.L., Nieto-Castanon, A.: Conn: A functional connectivity toolbox for correlated and anticorrelated brain networks. Brain Connectivity 2(3), 125–141 (2012)
12. Xu, P., Huang, R., Wang, J., Dam, N.T.V., Xie, T., Dong, Z., Chen, C., Gu, R., Zang, Y.F., He, Y., Fan, J., Luo, Y.: Different topological organization of human brain functional networks with eyes open versus eyes closed. NeuroImage 90 (2014)

Registering Cortical Surfaces Based on Whole-Brain Structural Connectivity and Continuous Connectivity Analysis

Boris Gutman, Cassandra Leonardo, Neda Jahanshad, Derrek Hibar, Kristian Eschenburg, Talia Nir, Julio Villalon, and Paul Thompson

Imaging Genetics Center, INI, University of Southern California, USA

Abstract. We present a framework for registering cortical surfaces based on tractography-informed structural connectivity. We define connectivity as a continuous kernel on the product space of the cortex, and develop a method for estimating this kernel from tractography fiber models. Next, we formulate the kernel registration problem, and present a means to non-linearly register two brains' continuous connectivity profiles. We apply theoretical results from operator theory to develop an algorithm for decomposing the connectome into its shared and individual components. Lastly, we extend two discrete connectivity measures to the continuous case, and apply our framework to 98 Alzheimer's patients and controls. Our measures show significant differences between the two groups.

Keywords: Diffusion MRI, Cortical Surface Registration, Connectivity Analysis, Data Fusion.

1 Introduction

With the advent of diffusion MRI, and the wealth of information contained within this modality, the subject of fusing structural connectivity information with anatomical knowledge has seen tremendous development. This fusion is straightforward if we restrict our diffusion analysis to summary voxel-wise measures such as Fractional Anisotropy (FA) or Mean Diffusivity. The problem becomes more difficult when we examine the connectivity information provided by tractography fiber models. Because fibers sets are not topologically equivalent across individual brains the usual solutions for image registration and segmentation problems cannot be trivially extended to these objects. Thus, it is not obvious how to fuse them with anatomical image processing in a straight-forward manner.

Several approaches have been proposed for fusing structural connectivity with anatomy. Perhaps the most common of these relies on the concept of a connectivity matrix or a graph between anatomically defined regions of interest (ROI). The strength of a connection between each region pair is estimated by counting the number of fiber models between the two ROI's [1]. The resulting graph can be analyzed using the standard graph theory measures [2], which can reveal interesting global and region-specific features of the brain's connectome, such as its "small-worldness," or the degree to which the network is compartmentalized into sub-networks [2].

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 161–168, 2014.

Alternatively, the DICCCOLs approach [3] seeks to identify small seed regions within the cortex which contain fibers with a similar geometric signature. The idea is that a geometric signature of the connection paths points to similar functional role across brains. Another exciting approach clusters brain regions spectrally with only the tractography seed regions as an anatomical prior [4].

Fiber- and anatomy-based registration fusion has also seen some development both with surface and volumetric anatomy models. Siless et al. [5] developed a framework based on geometric currents to drive inter-fiber set registration in combination with T1-weighted MRI image registration. Alternatively, Petrovic [6] assumed cortical alignment and registered thalamic surfaces based on the cortical fiber projections. In all of these cases, the full fiber geometry plays an integral part in driving the correspondence search, or some part of the brain is assumed to be perfectly aligned. The same is true for the region identification technique of DICCOLS: structural connectivity equivalence is estimated indirectly with a brief summary measure, defined as a histogram of orientations along the fiber. Unlike previous registration fusion approaches, we choose to apply the connectivity information supplied by the fiber model directly in a continuous registration setting, which significantly complicates the problem. Our goal is to find a correspondence between brains so that the corresponding regions are similarly connected. As in [1], we treat fiber geometry as a secondary feature, useful only in identifying the implied connection between brain regions. In this approach, two fibers with different geometry connecting the same pair of cortical locations are deemed equivalent.

Extending the discrete connectivity modeling of [2] to the continuous setting, we consider the connectome as a continuous kernel on the product space of the brain with itself. This is a natural extension of the graph representation for the discrete case. We treat each fiber as an instance of a connection on this space, with some possible geometric error. This idea naturally leads to kernel density estimation on the connectome space based on the set of fibers. Next, we would like to find a smooth non-linear invertible spatial warp that minimizes the difference between two brains' connectomes. Direct optimization of this problem poses a significant computational challenge. Instead, we decompose the kernel into its corresponding eigenfunctions, here called "eigen-networks," and use Mercer's Theorem for kernel matching and reconstruction. This convenient decomposition allows us to estimate the shared and subject-specific components of the connectome prior to registration, while the minimization problem is reduced to the usual multi-channel registration on the original domain of the cortex. We restrict our search to cortico-cortical and cortico-thalamic connections, which allows us to use the white matter boundary surface of the cortex to compactly represent the domain of the brain. Finally, we propose two continuous graph theory measures based on their discrete equivalents, and compute group differences between 48 Alzheimer's patients and 50 control participants from the ADNI cohort.

2 Continuous Connectome Estimation

We define the continuous connectome as a symmetric non-negative real-valued function $K: \mathcal{C} = \Omega \times \Omega \to \mathbb{R}^+$ by $(x, y) \mapsto K(x, y)$ from the product space of the cortical domain to the non-negative real numbers. $K(x, y)$ represents the strength of the

connection between the points x and y in the brain. As we are dealing with cortical surface models, our cortical domain is itself a mapping from the two-sphere into space: $\Omega = \mathcal{M}: \mathbb{S}^2 \to \mathbb{R}^3$. Since we perform our registration parametrically on $\mathbb{S}^2$, and because our $\mathcal{M}$ is diffeomorphic and area-preserving [7], we may equivalently set $\Omega = \mathbb{S}^2$ for convenience. While we do not have sufficiently resolved data to compute the true fiber-based connectivity, except using the coarsest resolution, we can apply the standard kernel density estimation. In this approach, we treat each fiber model as a representation of potentially many true fibers, with some possible error in its placement in the space $\mathcal{C}$. Given N_{fibers} fiber models, we project the two ends of each model onto the gray-white matter boundary, resulting in sets of point pairs $\{p_1^i, p_2^i\}$, discounting fibers that do not have both ends sufficiently close to the boundary. We apply the product of two Gaussian kernels on $\mathbb{S}^2$, [8] $G_\sigma: \mathbb{S}^2 \times \mathbb{S}^2 \to \mathbb{R}^+$, resulting in our non-local connectome estimation:

$$K_{non-local}(x,y) = \sum_{1=i\le N_{fibers}} G_\sigma(x, p_1^i)\, G_\sigma(p_2^i, y). \tag{1}$$

The parameter σ is set empirically so that the spherical area within the half-maximum of G_σ is equal to $\frac{2\pi}{N_{fibers}}$.

An aspect of brain connectivity which does not arise in the discrete approach is the modelling of local connections. Because tractography fiber models do not capture local connections at our cortical mesh resolution, we estimate local connectivity based on cortical geometry alone. We set local connectivity as

$$K_{local}(x,y) = G_\sigma(x,y). \tag{2}$$

A brief literature search [9] suggests we a golden ratio of local to global connectivity at $R_g = 1/3$. Thus, we set the complete connectivity kernel as

$$K = \frac{1}{N_{fibers}}\left(K_{non-local} + R_g K_{local}\frac{\|K_{non-local}\|}{\|K_{local}\|}\right), \tag{3}$$

where $\|K\| = \left(\iint_{\mathcal{C}} K(x,y)^2 dxdy\right)^{1/2}$.

3 Kernel Registration

Given two connectomes $K_1(x,y)$, $K_2(x,y)$, we assume that K_1 and K_2 differ from their mutual connectivity profile by a scale s, a smooth invertible warp $f: \Omega \to \Omega$ and an additive *individual* component:

$$K_{12,\ mutual}(x,y) = s_i\left[K_i\left(f_i[x], f_i[y]\right) - K_{i,\ indiv}(f_i[x], f_i[y])\right]. \tag{4}$$

For convenience, we set $f_1 = Id$. The kernel norm defined in the previous section suggests a cost function analogous to the L^2 fidelity in image registration:

$$\mathrm{C}(K_1, K_2, f) = \iint_{\mathcal{C}} [K_1(x,y) - K_2(f[x], f[y])]^2 \, dxdy \tag{5}$$

There are two issues with this formulation. First, while we have scaled the *full* kernels identically, we cannot know that their *mutual* connectomes will have the same scale. Second, a direct optimization of (5) is computationally expensive, as every point update requires full domain integration. Instead, we would like to estimate the mutual and individual components of the kernels prior to registration, while decoupling the two instances of $f[x]$ in (5). To this end, we decompose the kernels into the eigenfunctions, or "eigen-networks," of their linear operators, $Ae_i = \lambda_i e_i$, where $Af[y] = \int_\Omega K(x,y) f(x) dx$. According to Mercer's Theorem [10], we can reconstruct a symmetric positive definite (SPD) kernel by $K(x,y) = \sum_i \lambda_i e_i(x) e_i(y)$. This well-known result from operator theory provides an unexpected utility towards solving the kernel registration problem. To use it, we must only satisfy the SPD condition, which can be done by setting $K(x,x) = \int_\Omega K(x,y) dy$.

Since we assume that $K_{12,\ mutual}$ is itself SPD, we can make the assumption in (4) slightly stronger, asserting that the eigen-networks of $K_{12,\ mutual}$ and $K_{i,\ indiv}$ are orthogonal. On the other hand, because our non-linear correspondence search is local, we assume that some spatial overlap between the corresponding eigen-networks of $K_{12,\ mutual}$ and those of K_1 and K_2 must already exist. Note that a similar assumption is prevalent in standard non-linear registration algorithms. This allows us to estimate the mutual connectome by projecting the eigen-networks of the target connectome onto the corresponding invariant subspaces of the moving template connectome. We estimate the likelihood that for some small f, the transformed network of A_1, $f * e_i^1$, belongs to A_2 by

$$P[\{f * e_i^1\} \epsilon\ spectrum(A_2)] = \frac{\langle A_2 e_i^1, e_i^1 \rangle}{\left\| A_2 e_i^1 \right\|^2}, \tag{6}$$

where $\langle g, h \rangle = \int_\Omega gh$. For networks passing a threshold, we estimate their eigenvalue for A_2 as $\omega = \frac{\langle A_2 e_i^1, A_2 e_i^1 \rangle}{\langle A_2 e_i^1,\ e_i^1 \rangle}$, and project e_i^1 onto the invariant subspace of A_2 defined by ω. Via this process, we identify the set of mutual eigen-networks of the target and the moving template connectomes, $\boldsymbol{E_M} = \{e_{i_k}^1, e_{j_k}^2, w_k\}$.

A major difference between spectral decomposition of matrices and infinite-dimensional operators relates to eigenvalue multiplicity. In particular, it is possible to have non-isolated eigenvalues, and infinite-dimensional invariant subspaces. However, because our operator kernels are finites sum of weighted basis functions, we can say that the operators are finite-rank, and therefore necessarily compact [10]. This fact has a nice practical implication: the multiplicity of the eigenvalues is at most countable with the only possible limit point at 0. This means that any neighborhood $N_\epsilon(\omega), \omega > \epsilon$, contains a finite number of eigenvalues counting multiplicity, which makes step 1 in the following search feasible even in the true continuous case:

Algorithm 1 (mutual connectome estimation)
Given SPD kernels K_1, K_2, their corresponding operators A_1, A_2 and the ordered spectral decompositions $\{e_i^1, \lambda_i^1\}$, $\{e_j^2, \lambda_j^2\}$, set mutual networks $\boldsymbol{E_M} = \{e_{i_k}^1, e_{j_k}^2, w_k\} = NULL$. k = 1

For i=1:N
Compute $P = \frac{\langle A_2 e_i^1, e_i^1 \rangle}{\|A_2 e_i^1\|^2}$ (6)
  if(P > P_tol)
    1. Set $\omega = \frac{\|A_2 e_i^1\|^2}{\langle A_2 e_i^1, e_i^1 \rangle}$, set $\boldsymbol{E_i} = \left\{ e_j^2 \middle| 2\frac{|\lambda_j^2 - \omega|}{\lambda_j^2 + \omega} < \epsilon \right\}$
Project e_i^1 onto $span\{\boldsymbol{E_i}\}$, $e_{proj} = Proj_{span\{\boldsymbol{E_i}\}} e_i^1$
    if($\|e_{proj}\|$ > proj_tol)
      3. $e_{proj} \rightarrow e_{proj}/\|e_{proj}\|$
      4. $n = \max_j \{\langle e_{proj}, e_j^2 \rangle \mid e_j^2 \in \boldsymbol{E_i}\}$, set $e_n^2 = e_{proj}$
      5. Re-orthonormalize $\boldsymbol{E_i}$, starting with e_n^2
      6. Estimate new eigenvalues $\lambda_j^2 = \|A_2 e_j^2\|$, $e_j^2 \in \boldsymbol{E_i}$
      7. $w_k = \sqrt{\lambda_n^2 \lambda_i^1}$, $i_k = i$, $j_k = n$.
      8. Insert $(e_{i_k}^1, e_{j_k}^2, w_k)$ into $\boldsymbol{E_M}$, k = k + 1
    endif
  endif
end
Return $\boldsymbol{E_M}$

The mutual connectome can now be estimated as $K_{12,\ mutual}(x, y) = \sum_{e_i^1 \in \boldsymbol{E_M}} \lambda_i^1 e_i^1(x) e_i^1(y) e_i^1$. Finally, we define the mutual connectome mismatch cost:

$$\mathrm{C}(\boldsymbol{E_M}, f) = \int_\Omega \sum_{\left(e_{i_k}^1, e_{j_k}^2, w_k\right) \in \boldsymbol{E_M}} w_k \left| e_{i_k}^1(x) - e_{j_k}^2(f[x]) \right|^2 dx. \tag{7}$$

The solution of this functional is straightforward, and has been described elsewhere. It is worth noting that the gradient direction (7) is invariant to kernel scale, as it is based on normalized eigen-networks. As our parametric domain is $\mathbb{S}^2$, we use a recent spherical fluid registration algorithm [7], incorporating mean and Gaussian curvature mismatch in addition to (7). In this way we combine anatomical and connectivity information, registering brain connectivity structure directly across subjects.

4 Continuous Connectomics

Use of graph theory in brain connectivity studies has exploded in recent years; to this end, we contribute two weighted continuous analogues of the nodal degree and clustering coefficient measures [2]. Nodal degree, defined for discrete weighted graphs as

$k_d(x) = \sum_{y \in \Omega_d} K_d(x,y)$, where the subscript d means the discrete analogue of previous definitions, is defined here as the *connectedness map*

$$k(x) = \int_\Omega K(x,y)dy = K(x,x). \tag{8}$$

The discrete clustering coefficient is defined as $C_d(x) = \frac{2t_d(x)}{k_{db}(x)(k_{db}(x)-1)}$, where k_{db} is the binarisation of k_d, and $t_d(x) = \sum_{y,z \in \Omega_d} [K_d(x,y)K_d(x,z)K_d(y,z)]^{1/3}$ is the geometric mean of triangles. Our analogue geometric triangle mean is defined as

$$t(x) = \iint_{\mathcal{C}} [K(x,y)K(x,z)K(y,z)]^{1/3}\, dydz, \tag{9}$$

and the analogue of k_{db} is defined as the area of k's support:

$$C(x) = \frac{2t(x)}{\left[\int_{\{y \in \Omega | K(x,y)>0\}} dy\right]^2}. \tag{10}$$

5 Implementation

Our cortical surfaces are extracted with FreeSurfer, and mapped into correspondence on $\mathbb{S}^2$ by registering mean and Gaussian curvatures [7]. Tractography is performed by the Hough transform method [11], with fibers thresholded for length, resulting in 8-10K fibers. About 90% of these pass the threshold for interior ends being sufficiently close to the white matter surface, set at 10 mm. Connectome kernels are projected onto an equiangular spherical grid, with roughly 16.5K vertices per hemisphere, or 33K total. We use the Galerkin method [12] to estimate the eigen-network. The Galerkin method reduces an operator eigenvalue problem to a finite matrix problem, projecting the operator onto a finite set of basis functions. Our basis functions are the step functions defined by the equiangular sampling. We compute up to N eigenvalues, where N is the minimum number needed to approximate the kernel within a tolerance: $tol < \|K - K_N\|/\|K\|$, $K_N(x,y) = \sum_{i<N} \lambda_i e_i(x) e_i(y)$. We ignore the diagonal for this computation, as incorporating it gives optimistic error estimates.

We concede that a continuous formulation on paper often leads to the same implementation as a discrete one. In this case, though, the continuous formalism leads to a basic implementation difference: the area weights of the samples are taken into consideration. This is true both for the eigenvalue problem, which becomes generalized by the area matrix, and for continuous connectivity measures. In the latter case, we can think of the approximate kernel as a large weighted graph, *with weighted nodes*.

6 Experiments

We applied our method to 98 ADNI images. The participants were 48 AD patients and 50 controls. We chose an additional representative control subject to serve as the target. Following anatomical registration, we computed the connectome kernels and

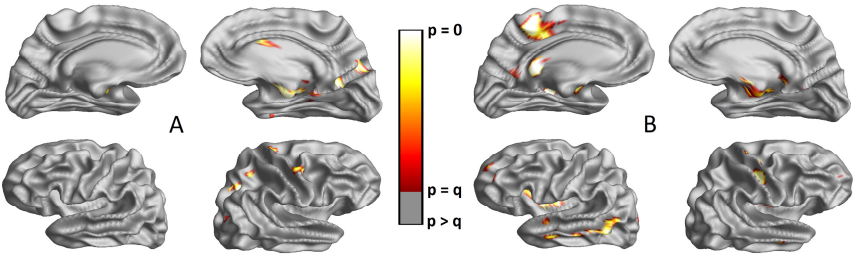


Fig. 1. Corrected p-maps for AD-NC difference in (A) clustering coefficient (10), and (B) connectedness, a.k.a. continuous nodal degree (8). Although it is mostly occluded, the left medial temporal lobe contains the most significant differences.

spectral decompositions of each subject. Each kernel had around 60M non-zero entries, making the connectome roughly 95% sparse. We set the kernel approximation tolerance at 0.1, requiring between 800 and 1200 eigen-networks. Approximately one-third of the target networks were matched to each moving kernel, depending on the participant. This set of networks was then registered to the target's while maintaining low curvature mismatch for anatomically correct correspondence, taking roughly 30 minutes for full combined connectivity registration.

In the first experiment, we computed the change in kernel mismatch, using both full and mutual network sets for kernel approximation. Results are displayed in Table 1, showing improvement in connectome alignment due to connectivity registration. In the second experiment, we performed a mass-univariate t-test over the cortical surface comparing connectedness and clustering coefficient maps between AD and control participants. Both measures passed False Discovery Rate (FDR) correction. FDR threshold for connectedness was $q = 1.0x10^{-3}$ for the right hemisphere, and $q = 1.9x10^{-3}$ for the left. For clustering coefficient, $q = 1.7x10^{-3}$ for the right hemisphere and $q = 2.5x10^{-5}$ for the left. In a related experiment, we made the same comparisons based only on anatomical registration. While the uncorrected p-maps were similar, right hemisphere connectedness and left clustering coefficient did not pass FDR. This suggests improved sensitivity due to the connectome registration. Corrected p-maps of these tests are displayed in Figure 1.

7 Conclusion

We have presented a framework for fusing connectivity information with cortical surface anatomy for a joint analysis. There are four distinct contributions: (1) the definition of a continuous connectome space and a method for estimating continuous kernels from fiber models; (2) an algorithm for defining a mutual connectome shared by two brains; (3) a spatial correspondence search between two connectome kernels, directly registering the brains' structural connectivities; (4) an adaptation of graph theory measures to the continuous setting. The final result is a pipeline for joint cortical surface and connectivity analysis that opens an exciting new way to explore the brain. Future work will ground our connectome estimation more strongly in biological knowledge and connect the eigen-network concept with functional connectivity.

Table 1. Relative difference between target and moving template connectomes before (Col. 1) and after (Col. 2) connectome registration (see section 5). Top row: full connectivity alignment. Bottom row: joint connectivity alignment. (Mean and standard deviation of 98 subjects).

	Anatomy only	Anatomy + connectivity	Individual improvement
Full kernel	0.528 +/-0.101	0.43 +/-0.12	0.098 +/-0.089
Mutual kernel	0.32 +/-0.082	0.12 +/-0.099	0.2 +/-0.089

References

1. Duarte-Carvajalino, J.M., Jahanshad, N., Lenglet, C., McMahon, K.L., de Zubicaray, G.I., Martin, N.G., Wright, M.J., Thompson, P.M., Sapiro, G.: Hierarchical topological network analysis of anatomical human brain connectivity and differences related to sex and kinship. Neuroimage 59, 3784–3804 (2012)
2. Rubinov, M., Sporns, O.: Complex network measures of brain connectivity: Uses and interpretations. Neuroimage 52, 1059–1069 (2010)
3. Zhu, D., Li, K., Guo, L., Jiang, X., Zhang, T., Zhang, D., Chen, H., Deng, F., Faraco, C., Jin, C., Wee, C.Y., Yuan, Y., Lv, P., Yin, Y., Hu, X., Duan, L., Han, J., Wang, L., Shen, D., Miller, L.S., Li, L., Liu, T.: DICCCOL: dense individualized and common connectivity-based cortical landmarks. Cereb Cortex 23, 786–800 (2013)
4. Lecoeur, J., Ingalhalikar, M., Verma, R.: Reproducibility of connectivity based parcellation: primary visual cortex. Proc. Int. Soc. Magn. Reson. Med. Sci. Meet. Exhib. Int. 2089 (2013)
5. Siless, V., Glaunès, J., Guevara, P., Mangin, J.-F., Poupon, C., Le Bihan, D., Thirion, B., Fillard, P.: Joint T1 and brain fiber log-demons registration using currents to model geometry. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part II. LNCS, vol. 7511, pp. 57–65. Springer, Heidelberg (2012)
6. Petrović, A., Smith, S., Menke, R., Jenkinson, M.: Methods for Tractography-Driven Surface Registration of Brain Structures. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009, Part I. LNCS, vol. 5761, pp. 705–712. Springer, Heidelberg (2009)
7. Gutman, B.A., Madsen, S.K., Toga, A.W., Thompson, P.M.: A Family of Fast Spherical Registration Algorithms for Cortical Shapes. Multimodal Brain Image Analysis (MBIA 2013) (2013)
8. Chung, M.K., Hartley, R., Dalton, K.M., Davidson, R.J.: Encoding Cortical Surface by Spherical Harmonics. Stat. Sinica 18, 1269–1291 (2008)
9. Rubenstein, J.L., Merzenich, M.M.: Model of autism: increased ratio of excitation/inhibition in key neural systems. Genes. Brain Behav. 2, 255–267 (2003)
10. Kreyszig, E.: Introductory functional analysis with applications. Krieger Pub. Co., Malabar (1989)
11. Aganj, I., Lenglet, C., Jahanshad, N., Yacoub, E., Harel, N., Thompson, P.M., Sapiro, G.: A Hough transform global probabilistic approach to multiple-subject diffusion MRI tractography. Medical Image Analysis 15, 414–425 (2011)
12. Beattie, C.: Galerkin Eigenvector Approximations. Math. Comput. 69, 1409–1434 (2000)

Automatic Method for Thalamus Parcellation Using Multi-modal Feature Classification

Joshua V. Stough[1], Jeffrey Glaister[2], Chuyang Ye[2], Sarah H. Ying[3], Jerry L. Prince[2], and Aaron Carass[2,4]

[1] Dept. of Computer Science, Washington and Lee University, Lexington, VA 24450, USA
[2] Dept. of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD 21218, USA
[3] Dept. of Radiology, The Johns Hopkins Hospital, Baltimore, MD 21287, USA
[4] Dept. of Computer Science, The Johns Hopkins University, Baltimore, MD 21218, USA

Abstract. Segmentation and parcellation of the thalamus is an important step in providing volumetric assessment of the impact of disease on brain structures. Conventionally, segmentation is carried out on T1-weighted magnetic resonance (MR) images and nuclear parcellation using diffusion weighted MR images. We present the first fully automatic method that incorporates both tissue contrasts and several derived features to first segment and then parcellate the thalamus. We incorporate fractional anisotrophy, fiber orientation from the 5D Knutsson representation of the principal eigenvectors, and connectivity between the thalamus and the cortical lobes, as features. Combining these multiple information sources allows us to identify discriminating dimensions and thus parcellate the thalamic nuclei. A hierarchical random forest framework with a multidimensional feature per voxel, first distinguishes thalamus from background, and then separates each group of thalamic nuclei. Using a leave one out cross-validation on 12 subjects we have a mean Dice score of 0.805 and 0.799 for the left and right thalami, respectively. We also report overlap for the thalamic nuclear groups.

Keywords: Brain imaging, diffusion MRI, magnetic resonance imaging, machine learning, segmentation, thalamus parcellation.

1 Introduction

The thalamus is a sub-cortical gray matter (GM) structure in the brain of vertebrates that is symmetric in the midline and located between the cerebral cortex and midbrain [18]. Its principal function is the relaying of sensory and motor signals to the cerebral cortex [18] and the regulation of consciousness, sleep, and alertness. The thalamus consists of lamellae—myelinated fibers—which separate the thalamus into its components and are grouped based on the orientation and location of distinct clusters of neurons. The most well known of these thalamic nuclear groups are the anterior nucleus (AN), medial dorsal (MD),

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 169–176, 2014.

ventral (VNG), pulvinar (PUL), lateral geniculate (LGN) and medial geniculate (MGN)—though each of these groups is made up of several smaller bundles of fibers. These nuclear groups are differentially affected in neurodegenerative diseases such as multiple sclerosis [8], Alzheimer's disease [4], schizophrenia [6,10], and Parkinson's disease [12]. Unfortunately, much of our understanding of the thalamus has come from neuropathological ex-vivo studies [6,10,12] which is not surprising considering that thalamic nuclei present minimal contrast in conventional MRI. Diffusion tensor imaging (DTI) presents a greater opportunity to unlock the secrets of the thalamus, as distinct tract connectivities and cytoarchitectures [16] provide a platform to distinguish the nuclear groups in-vivo. However, the exclusive use of DTI would make it impossible to distinguish the thalamus from other adjacent structures.

Previous work [11,14,17,20,23,24] has been limited to methods dependent on some level of manual interaction. This work presents two innovations: 1) it is the first fully automatic multi-modal thalamus segmentation algorithm, and 2) it is also the first fully automated thalamic nuclei parcellation—into AN, MD, VNG, PUL, LGN, and MGN—using tensor-based features within the thalamus and cortical connectivity features derived from tractography. Our method starts by generating an estimate of the region of interest (ROI) of the thalamus. Within this ROI, features are computed, including diffusion tensors and their principal directions and probabilistic connectivities between each voxel and lobar labels on the cerebral cortex. These features are used in a hierarchical random forest (RF) classifier framework, where the first RF segments the thalamus within the ROI, and a second RF identifies the collection of nuclear groups. The method is tested against manual delineation and its two phases (thalamic segmentation and nuclear group identification) are compared to other methods.

2 Method

2.1 ROI Identification

To reduce the computational burden of training an RF we estimate bounding boxes for the left and right thalami, denoted $\mathcal{B}_{\text{L}}$ and $\mathcal{B}_{\text{R}}$ respectively. These ROIs are identified using a tissue segmentation and labeling approach based on topology preservation and fuzzy classification [1]. For voxel j with spatial position $\mathbf{x}_j$ in the image domain Ω and with MR intensity $\mathcal{I}_j$, there are functions u_{jk} which represent the membership of the voxel with respect to structure k. The structures k have an intensity centroid of c_k. We introduce r_{jk} as a penalty term that discourages unrealistic configurations such as the thalamus touching the cerebellum. We have prior probabilities p_{jk} coming from a statistical atlas and weights w_{km} on the intensity difference between the centroids of two classes c_k and c_m. These terms are combined to form the following energy minimization problem,

$$E\left(u_{jk}\right) = \sum_{jk} \frac{u_{jk}^q}{r_{jk}} \left|\left|\mathcal{I}_j - c_k\right|\right|^2 + \beta \sum_{l \in N_j,\, m \neq k} \frac{u_{jk}^q}{r_{jk}} u_{lm}^q + \gamma \sum_{m \neq k} \frac{u_{jk}^q}{r_{jk}} w_{km} p_{jm}^q, \quad (1)$$

where q is a fuzziness parameter. The first term, on the right hand side of Equation (1) ensures voxels in the same structure have similar intensity values, while the second term controls the smoothness of the memberships, and the final term regulates the influence of the prior probability. β and γ are weights that balance the relative influence of the terms. The energy is minimized while simultaneously maintaining the topological arrangements of the objects achieved through max membership assignment.

Given a fuzzy segmentation estimate of the left thalamus $\mathcal{T}_{\text{L}}$, $\mathcal{B}_{\text{L}}$ is defined as

$$\mathcal{B}_{\text{L}} = \{\mathbf{x}_j | (\mathbf{l}_{\text{L}} - \mathbf{r}_{\text{L}}) \leq \mathbf{x}_j \leq (\mathbf{h}_{\text{L}} + \mathbf{r}_{\text{L}}), \mathbf{x}_j \in \Omega\} \tag{2}$$

where $\mathbf{l}_{\text{L}} = \arg\min_{\mathbf{x}_j \in \mathcal{T}_{\text{L}}} \mathbf{x}_j$, $\mathbf{h}_{\text{L}} = \arg\max_{\mathbf{x}_j \in \mathcal{T}_{\text{L}}} \mathbf{x}_j$, and $\mathbf{r}_{\text{L}} = 0.1 \times (\mathbf{h}_{\text{L}} - \mathbf{l}_{\text{L}})$, which pads $\mathcal{T}_{\text{L}}$ by 10% along each axes. This process is repeated for $\mathcal{B}_{\text{R}}$ from its corresponding fuzzy segmentation $\mathcal{T}_{\text{R}}$. Henceforth, when we refer to $\mathcal{B}$ it is implied that the process is repeated for both $\mathcal{B}_{\text{L}}$ and $\mathcal{B}_{\text{R}}$, independently.

2.2 Knutsson Space and Edge Maps

DTI is acquired from diffusion weighted MRI, using a gradient spin echo pulse sequence with a known b-value b and gradient direction $\mathbf{g}$. The diffusion signal, $S(b, \mathbf{g})$, at each voxel is an attenuated version of the signal S_0 that would be recorded in the absence of diffusion weighting. The relationship can be specified using the Stejskal-Tanner equation,

$$S(b, \mathbf{g}) = S_0 e^{-b\mathbf{g}^T D \mathbf{g}} \tag{3}$$

where D is the 3×3 symmetric diffusion tensor,

$$D = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{xy} & D_{yy} & D_{yz} \\ D_{xz} & D_{yz} & D_{zz} \end{bmatrix} = [\mathbf{u}_1\, \mathbf{u}_2\, \mathbf{u}_3] \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} [\mathbf{u}_1\, \mathbf{u}_2\, \mathbf{u}_3]^T . \tag{4}$$

The eigenvalues $(\lambda_1, \lambda_2, \lambda_3)$ from Equation (4) have eigenvectors $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$.

Two common quantities computed from the eigenvalues are the mean diffusivity (MD) and fractional anisotrophy (FA), denoted $\mathcal{M}$ and $\mathcal{F}$, respectively. The principal eigenvector (PEV) $\mathbf{u}_1$ represents the direction of maximum diffusion. As the diffusion occurs either in the direction of $\mathbf{u}_1$ or in the opposite direction $-\mathbf{u}_1$ with equal probability, it is convenient to represent the direction $\mathbf{u}$ as an *orientation* using the Knutsson map [15], which transforms the eigenvector $\mathbf{u}$ from $\mathbb{S}^2$ to $\mathbb{K} \subset \mathbb{R}^5$ by

$$K(\mathbf{u} = (u_1, u_2, u_3)) = \left(u_1^2 - u_2^2,\, 2u_1u_2,\, 2u_1u_3,\, 2u_2u_3,\, \frac{1}{\sqrt{3}} \left(2u_3^2 - u_1^2 - u_2^2\right) \right). \tag{5}$$

This mapping takes opposing Cartesian vectors and sends them to the same location in Knutsson space $\mathbb{K}$—that is both length and direction are crushed in the transformation to $\mathbb{K}$. We can now generate an edge map using orientations

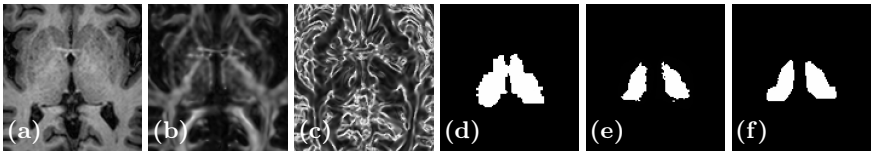


Fig. 1. Shown are **(a)** the MPRAGE $\mathcal{I}$, **(b)** the FA $\mathcal{F}$, and **(c)** the edgemap $\|G\|_F$. Thalami estimates from **(d)** FreeSurfer [9], **(e)** our method (OM 18F), and **(f)** a manual delineation.

in $\mathbb{K}$. For $\mathbf{v} = (v_1, \ldots, v_5) \in \mathbb{K}$ we have the gradient matrix G and its Frobenius norm $\|G\|_F$ given by

$$G(\mathbf{v}) = \begin{bmatrix} \frac{\partial v_1}{\partial x} & \frac{\partial v_1}{\partial y} & \frac{\partial v_1}{\partial z} \\ \vdots & \vdots & \vdots \\ \frac{\partial v_5}{\partial x} & \frac{\partial v_5}{\partial y} & \frac{\partial v_5}{\partial z} \end{bmatrix} \qquad \|G(\mathbf{v})\|_F = \sqrt{\sum_j \sum_i G_{ij}^2}\,. \tag{6}$$

This is an edge map representing a change in the direction of the PEV, which will allow us to distinguish thalamic nuclei.

2.3 Connectivity to the Cortical Mantle

Connectivity to the cortical mantle is calculated using probabilistic tractography [3] implemented in the FSL toolkit. Six cortical masks corresponding to thalamic connection sites are used as the targets for the tractography algorithm. M_l is the set of voxels in a cortical mask and l is the cortical mask label. The six labels for the cortical masks are {frontal, occipital, parietal, temporal, precentral, postcentral}.

Connectivity $C_l(\mathbf{x})$ is defined as the number of times a sample starting at the voxel $\mathbf{x}$ forms a pathway connecting to any voxel $\mathbf{y}$ belonging to the cortical mask with label l,

$$C_l(\mathbf{x}) = \frac{|\{\mathbf{y} | \exists \mathbf{x} \rightarrow \mathbf{y} \in M_l\}|}{|M_l|}\,. \tag{7}$$

5000 samples are initiated per voxel in $\mathcal{B}$ and the path direction is determined by local fiber directions.

2.4 Features and Random Forest Framework

The first features input into our RF framework are the relative position of $\mathbf{x}_j \in \mathcal{B}$ and the MR intensity value at $\mathbf{x}_j$, $\mathcal{I}_j$. These intensities provide clues about the boundary of the thalamus with non-thalamus structures. The core distinguishing features of the nuclear groups are fiber orientation and strength. Thus the next set of features are the FA, MD, Knutsson mapping, and Frobenius norm, denoted as $\mathcal{F}_j, \mathcal{M}_j, \{K(\mathbf{x}_j)\}$, and $\|G(K(\mathbf{x}_j))\|_F$, respectively. The final features are the

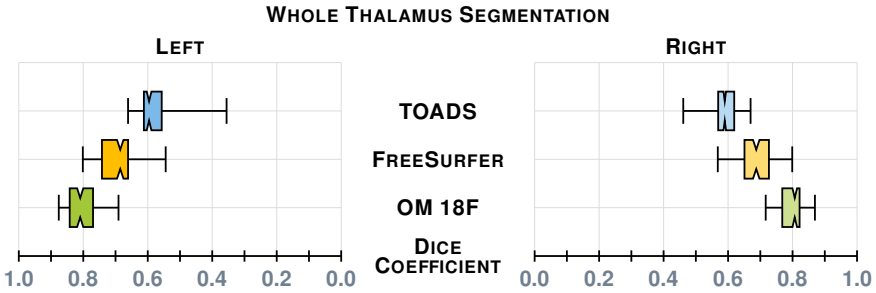


Fig. 2. A comparison between our thalamus segmentation (OM 18F) and those of Bazin and Pham [1] (TOADS) and Dale et al. [9] (FreeSurfer 5.3.0). The notches give a 95% confidence interval for the difference in two medians.

connectivity between the position $\mathbf{x}_j$ and the six cortical labels (i.e., the $C_l(\mathbf{x}_j)$'s), denoted $\{C(\mathbf{x}_j)\}$. The complete feature vector $\mathbf{f}_j$ is

$$\mathbf{f}_j = (\mathbf{x}_j, \mathcal{I}_j, \mathcal{F}_j, \mathcal{M}_j, \{K(\mathbf{x}_j)\}, \|G(K(\mathbf{x}_j))\|_F, \{C(\mathbf{x}_j)\}) \tag{8}$$

which gives us an 18-dimensional feature space.

Our hierarchical RF [5] approach uses $\mathbf{f}_j$ for each voxel in the available training data to build a collection of trees that first distinguishes the thalamus within $\mathcal{B}$ from other tissues. This is a binary classification task identifying thalamus from background. A second RF is then built using the same feature vector, trained to provide a membership for each of the six thalamic nuclear groups given that we know the thalamus from the first stage. The first stage thalamus identification can be quite noisy due to peripheral objects have a thalamus-like appearance. To reduce this artifact, we select the largest connected component foreground object which we then close with a $3 \times 3 \times 3$ structuring element. The learnt RFs can be applied to a new subject, with the classification scores determining the segmentation of the thalamus and subsequent parcellation of the nuclear groups.

3 Results

3.1 Data

Our data consists of 12 subjects from a study of cerebellar ataxia. The subject images were acquired on a 3T MR scanner (Intera, Philips Medical Systems, Netherlands) and have undergone standard neuroimaging processing: inhomogeneity correction [19], skull stripping [7], isotropic resampling [22] to 0.828 mm, distortion correction [21], and probabilistic tractography [13]. A subject is shown in Fig. 1 showing some of the input contrasts. We refer to our method as OM 18F, as in our method using 18 features.

A manual rater first used the FA to find the thalamus boundary, then used the Knutsson edge map to delineate nuclear structures that we identify as the AN,

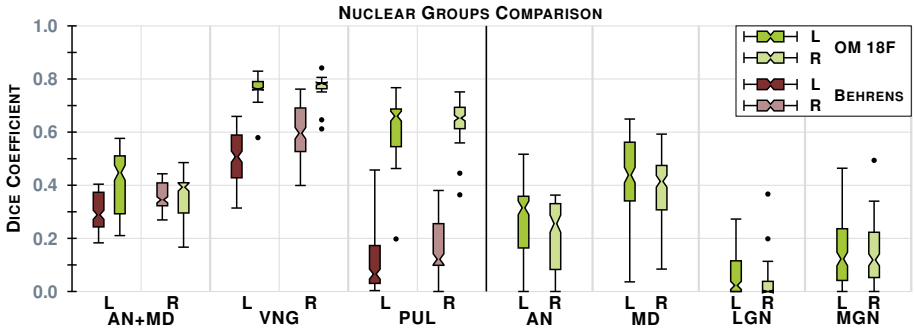


Fig. 3. The plot on the left is a comparison of the Dice score between our implementation of Behrens et al. [2] (B) (shades of brown) and our method (OM 18F) (shades of green), see the text for details. On the right is the remaining nuclear groups we can generate. Both plots are for our 12 subjects. Results for left and right thalami are denoted **L** and **R**, respectively. The notches give a 95% confidence interval for the difference in two medians.

MD, PUL, LGN, and MGN nuclei. VG is the complement of these structures, within the thalamus boundary. We use these reproducible manual delineations as a ground truth for our training and testing.

3.2 Thalamus Boundary

Our first results compare our estimate of the thalamus with those from two whole brain segmentation software tools [1,9]. We used leave-one-out cross-validation to train both our RFs, the results are averaged over the different cross-validation runs and Dice scores are shown in Fig. 2. A paired Wilcoxon rank sum test comparing our method with Bazin and Pham [1] (TOADS) had a p-value < 0.001 for both the left and right thalami (computed independently), indicating significant improvement. A similar test between our method and Dale et al. [9] (FreeSurfer) gives a p-value < 0.001 for the right thalamus; however for the left thalamus the p-value is 0.00684, which is just shy of statistical improvement. We note that in this stage, as in the next, there are two RFs one for the left thalamus and the other for the right stemming from $\mathcal{B}_L$ and $\mathcal{B}_R$, respectively. Example results and comparison to our ground truth is shown in Fig. 1.

3.3 Thalamic Nuclei Segmentation

The second step in our hierarchical RF framework distinguishes thalamic nuclei assuming that the thalamus boundary is known from the first stage. The left and right thalami were those identified in Section 3.2, which were passed through their respective trained RFs to predict the nuclear groups. We also implemented an automated algorithm based on Behrens et al. [2], which used only the cortical labels to parcellate the thalamus. The fiber groupings in Behrens et al. [2] are

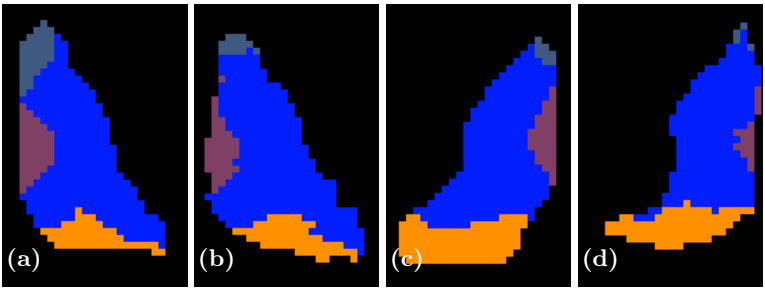


Fig. 4. Shown are axial slices of **(a)** a manual delineation and **(b)** our parcellation for a right thalamus on one of our better results and **(c)** a manual delineation and **(d)** our parcellation of a left thalamus for a bad result. The AN is shown in a slate blue anterior to the thalamus, the VNG is the large blue body in the center of the thalamus, while MD and PUL are shown in purple and orange, respectively.

different to ours, thus for comparison we merged AN & MD (AN+MD). Behrens et al. [2] also excluded LGN and MGN from their parcellation. The comparable nuclear groups—AN+MD; VNG; PUL—are shown in Fig. 3, as well as the additional nuclear groups we can parcellate. A paired Wilcoxon rank sum test comparing the results has a p-value < 0.001 for the VNG and PUL on both thalami. We fail to reach significance when comparing for AN+MD, because we train for AN and MD separately; training on the merging of these groups would perform better. Examples of our parcellation for two subjects are shown in Fig. 4.

4 Conclusion

In this paper we have presented the first fully automatic thalamic parcellation method using multi-modal imaging data, and we make two important contributions. Firstly we use a multi-channel framework to segment the thalamus—the first such method. Secondly, we provide a parcellation of the six core nuclear groups of the thalamus in a fully automated fashion.

Acknowledgments. This work was supported by the NIH/NINDS grants R21-NS082891 and R01-NS056307. Jeffrey Glaister is supported by the Natural Sciences and Engineering Research Council of Canada. We would like to thank the MICCAI reviewers for their suggestions in helping improve our approach.

References

1. Bazin, P.L., Pham, D.L.: Homeomorphic brain image segmentation with topological and statistical atlases. Medical Image Analysis 12(5), 616–625 (2008)
2. Behrens, T.E.J., et al.: Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. Nature Neuroscience 6(7), 750–757 (2003)

3. Behrens, T.E.J., et al.: Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? NeuroImage 34(1), 144–155 (2007)
4. Braak, H., Braak, E.: Alzheimer's disease affects limbic nuclei of the thalamus. Acta Neuropathologica 81(3), 261–268 (1991)
5. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
6. Byne, W., et al.: Postmortem Assessment of Thalamic Nuclear Volumes in Subjects With Schizophrenia. Am. J. Psychiatry 159(1), 59–65 (2002)
7. Carass, A., et al.: Simple paradigm for extra-cerebral tissue removal: Algorithm and analysis. NeuroImage 56(4), 1982–1992 (2011)
8. Cifelli, A., et al.: Thalamic neurodegeneration in multiple sclerosis. Annals of Neurology 52(5), 650–653 (2002)
9. Dale, A.M., et al.: Cortical Surface-Based Analysis I: Segmentation and Surface Reconstruction. NeuroImage 9(2), 179–194 (1999)
10. Danos, P., et al.: Volumes of association thalamic nuclei in schizophrenia: a postmortem study. Schizophrenia Research 60(2-3), 141–155 (2003)
11. Duan, Y., et al.: Thalamus Segmentation from Diffusion Tensor Magnetic Resonance Imaging. International Journal of Biomedical Imaging 5(2), 1–5 (2007)
12. Jellinger, K.A.: Post mortem studies in Parkinson's disease–is it possible to detect brain areas for specific symptoms? J. Neural. Transm. Suppl. 56, 1–29 (1999)
13. Jenkinson, M., et al.: FSL. NeuroImage 62(2), 782–790 (2012)
14. Jonasson, L., et al.: A level set method for segmentation of the thalamus and its nuclei in DT-MRI. Signal Processing 87(2), 309–321 (2007)
15. Knutsson, H.: Producing a Continuous and Distance Preserving 5-D Vector Representation of 3-D Orientation. In: IEEE Computer Society Workshop on Computer Architecture for Pattern Analysis and Image Database Management, pp. 175–182 (1985)
16. Morel, A., et al.: Multiarchitectonic and Stereotactic Atlas of the Human Thalamus. J. Comparative Neurology 387(4), 588–630 (1997)
17. Rittner, L., et al.: Segmentation of thalamic nuclei based on tensorial morphological gradient of diffusion tensor fields. In: 7th International Symposium on Biomedical Imaging (ISBI 2010), pp. 1173–1176 (2010)
18. Sherman, S.M., Guillery, R.: Exploring the Thalamus. Elsevier (2000)
19. Sled, J.G., et al.: A non-parametric method for automatic correction of intensity non-uniformity in MRI data. IEEE Trans. Med. Imag. 17(1), 87–97 (1998)
20. Stough, J., et al.: Thalamic Parcellation from Multi-Modal Data using Random Forest Learning. In: 10th International Symposium on Biomedical Imaging (ISBI 2013), pp. 852–855 (2013)
21. Studholme, C., et al.: Accurate alignment of functional EPI data to anatomical MRI using a physics-based distortion model. IEEE Trans. Med. Imag. 19(11), 1115–1127 (2000)
22. Thévenaz, P., et al.: Interpolation revisited (medical images application). IEEE Trans. Med. Imag. 19(7), 739–758 (2000)
23. Wiegell, M.R., et al.: Automatic segmentation of thalamic nuclei from diffusion tensor magnetic resonance imaging. NeuroImage 19(2), 391–401 (2003)
24. Ye, C., et al.: Parcellation of the Thalamus Using Diffusion Tensor Images and a Multi-object Geometric Deformable Model. In: Proceedings of SPIE Medical Imaging (SPIE-MI 2013), Orlando, FL, February 9-14, vol. 8669, pp. 866909–866909–7 (2013)

Multiple-Network Classification of Childhood Autism Using Functional Connectivity Dynamics

True Price, Chong-Yaw Wee, Wei Gao, and Dinggang Shen

Department of Radiology and Biomedical Research Imaging Center (BRIC)
The University of North Carolina at Chapel Hill, USA
jtprice@cs.unc.edu, {cywee,dinggang_shen}@med.unc.edu,
wgao@email.unc.edu

Abstract. Characterization of disease using stationary resting-state functional connectivity (FC) has provided important hallmarks of abnormal brain activation in many domains. Recent studies of resting-state functional magnetic resonance imaging (fMRI), however, suggest there is a considerable amount of additional knowledge to be gained by investigating the variability in FC over the course of a scan. While a few studies have begun to explore the properties of dynamic FC for characterizing disease, the analysis of dynamic FC over multiple networks at multiple time scales has yet to be fully examined. In this study, we combine dynamic connectivity features in a multi-network, multi-scale approach to evaluate the method's potential in better classifying childhood autism. Specifically, from a set of group-level intrinsic connectivity networks (ICNs), we use sliding window correlations to compute intra-network connectivity on the subject level. We derive dynamic FC features for all ICNs over a large range of window sizes and then use a multiple kernel support vector machine (MK-SVM) model to combine a subset of these features for classification. We compare the performance our multi-network, dynamic approach to the best results obtained from single-network dynamic FC features and those obtained from both single- and multi-network static FC features. Our experiments show that integrating multiple networks on different dynamic scales has a clear superiority over these existing methods.

1 Introduction

Resting-state functional connectivity (FC) has been proven to be a critical tool in understanding different disease mechanisms and has great potential to provide biomarkers for disease diagnosis and monitoring [6]. Canonical models of altered connectivity among specific regions of the brain have been proposed for a wide range of neurological diseases, including Alzheimer's Disease [8], schizophrenia [7], and autism [12]. In the past, many of these functional characterizations of mental disease have assumed that connectivity patterns in the brain do not change over the course of a resting-state fMRI scan. There is a growing consensus in the neuroimaging community, however, that FC fluctuates in a task-free environment with correspondence to cognitive state [2,9]. These short-scale modulations in connectivity, which were latent under previous assumptions of stationarity FC, accordingly contain valuable information about functional

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 177–184, 2014.

organization in the resting-state brain. Utilizing the novel features of FC dynamics, in turn, may allow us to build a better understanding of the effects of neurological disease on brain function.

It is only recently that dynamic FC patterns have been investigated for the purpose of characterizing disease. Ma *et al.* [14] defined group-level intrinsic connectivity networks (ICNs) in the resting-state brain and assessed differences in schizophrenic dynamic FC patterns using a sliding windows approach. Additionally, they used Markov modeling to identify abnormal modulation of brain states within the default mode network (DMN) of the schizophrenic group. Leonardi *et al.* [13] used principal components analysis over sliding windows to uncover differences in whole-brain FC variation patterns (termed "eigenconnectivites") between relapse-remitting multiple sclerosis (RRMS) patients and normal controls. This novel approach revealed altered DMN-related eigenconnectivities among RRMS patients.

In addition to dynamic changes in connectivity on a single network scale, the interaction of multiple large-scale networks has recently become an important topic in brain disease investigation [5,11]. As the brain's functioning is a product of many concurrent neural patterns, the incorporation of multiple networks into resting-state analysis has great bearing for robust disease characterization. To date, however, there has been little research addressing the properties of dynamic connectivity in disease on a multi-network scale.

In this paper, we combine the dynamic properties of functional connectivity among multiple resting-state networks and explore the merits of such an approach in better classifying childhood autism spectrum disorders (ASDs). Using ICNs defined from group independent component analysis (ICA) on resting-state fMRI data, we investigate sliding window connectivity within individual networks on a large range of time scales. We then apply a multiple kernel support vector machine (MK-SVM) model to evaluate the combination of multiple networks on multiple scales and compare the classification results to those obtained from single-network analysis or under the assumption of stationary functional connectivity over the course of the scan.

2 Methods

Our approach to ASD classification uses multi-network combination of intra-network dynamic connectivity features. Starting with ICNs defined at the group level, we employed linear regression to recover subject-specific time courses for each network. Then, dynamic FC features were extracted for each network using sliding window correlations over a large range of window sizes (20 to 240 seconds, at step sizes of 10 seconds); stationary connectivity features were also collected. Using this wide array of feature types over all networks, we applied an iterative selection/weighting algorithm in a multiple kernel SVM (MK-SVM) model to identify optimal ICNs and window sizes for overall ASD classification. Leave-one-out cross-validation (LOOCV) was used to separate the data into the training and testing sets used in classification. An overview of the classification pipeline is illustrated in Figure 1.

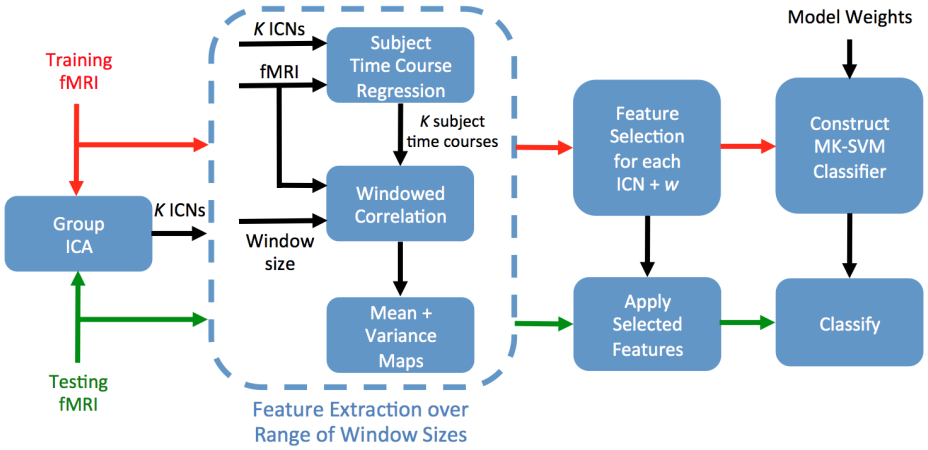


Fig. 1. Framework for the proposed classification pipeline. Subject-specific dynamic connectivity features are derived from group-level ICNs on multiple time scales (i.e. sliding window sizes). Feature selection is then applied to each ICN for every time scale. Finally, features from all ICNs at all time scales are combined in a weighted multiple kernel model.

2.1 Participant Data

Resting-state subject scans were obtained from the open-access Autism Brain Imaging Data Exchange (ABIDE) database [4]. A cohort of 60 child scans, 30 categorized as typical controls (TC) and 30 diagnosed with ASD, were selected from the NYU Langone Medical Center ABIDE site dataset. Mean group ages, in years, were 9.75 ± 1.40 for ASD patients and 9.69 ± 1.58 for controls. Subject ages ranged between 6.5 and 12 years, and individuals were selected to minimize between-group age differences (p = 0.8873). Information about participant data collection, exclusion criteria, and scan parameters for the NYU dataset is available on the ABIDE website[1].

2.2 Data Preprocessing

Initial fMRI scans were collected on a 3-Tesla Siemens Allegra scanner over six minutes taking 180 time points at a repetition time (TR) of 2s. The data were preprocessed using Data Processing Assistant for Resting-State fMRI (DPARSF) software [3]. Before preprocessing, all images had the first ten time points removed. The remaining volumes were then normalized to MNI space with a resolution of $3 \times 3 \times 3$ mm^3. Next, the images were slice timing corrected and motion corrected using the first remaining time point as a reference. White matter, CSF, global signals, and head motion were regressed out as nuisance covariates. Following this, the images underwent signal detrending and band-pass filtering (0.01-0.08Hz). Finally, motion scrubbing [16] was applied with an FD threshold of 0.5; time points with significant motion were removed from each image, along with the preceding time point and the two time points following.

[1] http://fcon_1000.projects.nitrc.org/indi/abide/

2.3 Group ICA and Recovery of Subject Time Courses

In this study, we employed group ICA to define population-based ICNs. Using FSL's MELODIC software[2] [17], all preprocessed subject images – both ASD and normal controls – were temporally concatenated and projected into a 25-dimensional subspace. Spatial ICA was performed on this data set to recover 25 statistically independent spatial maps, each representing a unique group-level functional network, with associated group-level time-courses. After recovering group-level ICNs, we followed the first step of FSL's dual regression approach [1] to define subject-specific time courses associated with each individual network. Namely, we performed linear regression to model each time point of an individual's fMRI scan as a linear sum of the group-level spatial maps.

2.4 Estimating Intra-network Functional Connectivity

For each subject, the back-reconstruction process yielded a set of 25 time courses, each representing an underlying signal associated with a single group-level functional network. Then, for each subject, we measured the influence of a functional network i on a given voxel v as the normalized cross-correlation of their respective time courses:

$$I(i,v) = \frac{1}{N}\sum_{t=1}^{N}\frac{(T_i(t)-\overline{T_i})(T_v(t)-\overline{T_v})}{\sigma_i\sigma_v}, \tag{1}$$

where N is the length of both time courses, T_i is the subject-specific time course representing network i, T_v is the BOLD signal of voxel v, and $\overline{T}$ and σ represent the mean and standard deviation of a time course, respectively. To analyze the change in network influence over the course of the scan, we split T_i and T_v into synchronous sliding time windows and computed the correlation for each window separately. To observe the effect of time scale on dynamic functional connectivity within the networks, we repeated the experiment on a large range of window sizes, from 10 TR (20 s) to 120 TR (240 s) at intervals of 5 TR. For all window sizes, we fixed the step size between windows to 2 TR (4 s).

Because resting-state fMRI is inherently task-free, it is difficult to interpret patterns of dynamic FC in individuals, and likewise, it is impractical to directly compare changes in FC between subjects. To account for this, we took the voxel-wise mean and variance of windowed correlation values for a given window size and used these measures as comparative features. For a given voxel and a given network, the mean correlation over all time windows gives the average influence of the network within that voxel on a specific dynamic scale. Similarly, the variance gives the local stationarity of the network influence on the same scale. That is, voxels with high variance in correlation can be interpreted to experience some shift in intra-network functional connectivity over the course of the scan.

2.5 Feature Selection

As ICA may return artifactual or physiological components, we first visually inspected the set of 25 components and selected 16 as relevant to functional dynamics, discarding

[2] `http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/MELODIC`

the rest [10]. Then, prior to classification, we performed the following feature selection steps for each ICN: First, we masked each ICN to only include functionally relevant voxels, as determined by MELODIC's post-processing mixture model [17] (using a p-value cutoff of $p < 0.0001$). Next, we applied the Mann-Whitney test [15] to only select mean and variance features where one group exhibited significantly higher values than the other group (thresholded at $p < 0.05$). Using the remaining features, we further applied a logistic regression with L_1-norm regularization to ensure a small number of features were used when constructing linear kernels, which can be strongly affected by noisy features, in our MK-SVM model. Importantly, feature selection, as well as subsequent classifier training, was performed independently on each training set in the LOOCV framework.

2.6 Multiple Network Classification with MK-SVM

From our collection of ICNs, each evaluated at a range of window sizes, our next step is to integrate the features from all networks to perform combined classification. In practice, we expect that only a subset of ICNs actually contain meaningful differences in autism connectivity, and that certain window sizes will provide better disease discriminability than others for ICNs that do have important features. Therefore, we wish to develop a minimal, yet multi-network, multi-scale model of ICN connectivity that increases ASD classification over both single-network analysis and assumptions of FC stationarity. To evaluate the feasibility of such a design, we propose to use an iterative ICN selection/weighting method to find the combination of ICN features that maximizes overall LOOCV classification accuracy, using a multiple kernel SVM (MK-SVM) model [20] for classification.

Given a set of K kernels $\{\phi_k(\mathbf{x})\}$ generated from the training set $\{(\mathbf{x}_i, y_i)\}$, where $\mathbf{x}_i \in \mathbb{R}^{N\times 1}$ is a feature vector and $y_i \in \{-1, 1\}$ is a class label, MK-SVM seeks to find a maximum margin hyperplane in kernel space to separate the two classes. The primal formulation of MK-SVM seeks to solve

$$\min_{\mathbf{w}_k, b, \xi_i} \frac{1}{2}\sum_{k=1}^{K} \beta_k ||\mathbf{w}_k||^2 + C\sum_{i=1}^{N} \xi_i \tag{2}$$

$$\text{s.t. } \xi \geq 0, y_i \left(\sum_{k=1}^{K} \beta_k \left(\mathbf{w}_k^T \phi_k\left(\mathbf{x}_i\right) + b\right)\right) \geq 1 - \xi_i \text{ for } i = 1, \ldots, n.$$

Here, β_k is a linear weight for kernel k; $\mathbf{w}_k$ and b are the normal vector and intercept defining the hyperplane, respectively; and C is a parameter regularizing the degree of misclassification (we used a default value of $C = 1$ in our experiments). Given a test observation $\mathbf{x}$, we can then make a prediction on its class $\hat{y}$ to be

$$\hat{y} = sign\left(\sum_{k=1}^{K} \beta_k \left(\mathbf{w}_k^T \phi_k\left(\mathbf{x}\right) + b\right)\right). \tag{3}$$

In practice, we can constrain the kernel weights such that $\sum_{k=1}^{K} \beta_k = 1$ and perform a coarse search within the combinatorial space to select an optimal weight configuration.

Since we only wish to select a subset of ICNs out of a large set of network/window size combinations, we employed a greedy forward approach to select kernels one-at-a-time while simultaneously learning their associated MK-SVM weights $\{\beta_k\}$. In this approach, we first construct a linear kernel for each ICN and for each window size, forming the set of kernels $\{\phi_k(\mathbf{x})\}$. We then seek to find a subselection of K kernels that maximizes LOOCV accuracy in the MK-SVM model. Starting with the ICN that gives the best classification accuracy, we incrementally add an additional ICN to the model if its inclusion increases overall accuracy. At the same time, we use a grid search approach to select the optimal weights of selected kernels in the MK-SVM model. For each iteration, we generate all combinations of kernel weights between 0 and 1, using a step size equal to the inverse of the number of kernels in the model. We stop our iterative addition of ICNs when overall classification accuracy can no longer be improved.

3 Results and Discussion

Table 1 gives the best results of dynamic and static FC classification on both single- and multi-network approaches. For the single-network case, we report the highest LOOCV accuracy among all 16 networks using each network's features as input to a single, linear kernel SVM. Under assumptions of FC stationarity, the maximum single-network accuracy was 68%; in our experiments, this network corresponds to the salience network, which under static FC analysis has previously been shown to have a similar rate of disease discriminability children with ASD [19]. For dynamic FC analysis, we found the central executive network (CEN) to have a classification rate of 83% at a window size of 35 TR (70 s). Interestingly, it has been suggested that disruptions in high-order cognitive switching controlled by the CEN could contribute to several neurophysiological disorders, including autism [18]. Figure 2 (left) shows the salience and central executive networks recovered by ICA in our experiments. The right image in Figure 2 plots the classification performance of the two networks across all tested window sizes. We observe from the figure that the salience network has the best performance under static IC analysis, while the CEN has a peak between 30 and 40 TR. This finding is in line with our hypothesis that important FC information exists at different window sizes for different ICNs, and furthermore, it underscores the importance of considering multiple time scales when assessing dynamic FC features in mental disease.

Our results also suggest that multi-network classification, which is able to capture a larger range of disease characteristics, augments both stationary and dynamic analyses.

Table 1. Comparison of the best classification performance obtained under single- and multi-network models with and without dynamic FC features. (ACC=Accuracy, SEN=Sensitivity, SPE=Specificity, PPV=Positive Predictive Value, NPV=Negative Predictive Value)

Method	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)
Single-Network, Static FC	68	70	67	68	69
Single-Network, Dynamic FC	83	83	83	83	83
Multi-Network, Static FC	83	87	80	72	86
Multi-Network, Dynamic FC	90	87	93	93	88

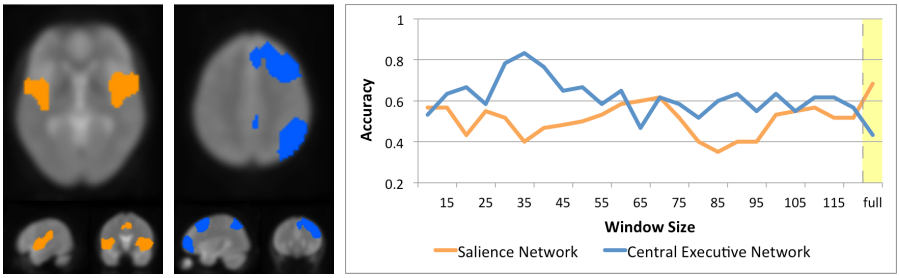


Fig. 2. Left: The salience and central executive networks, which displayed the highest classification accuracy among all 16 networks using static and dynamic connectivity features, respectively. Right: Classification accuracy of the two networks as a function of window size. Here, "full" denotes classification under stationary FC assumptions.

Moreover, compared to only using static FC features, we find that searching over multiple dynamic ranges enhances MK-SVM performance for the combination of ICNs. The multi-network, multi-scale model obtained 90% accuracy using the approach adopted here, substantially outperforming the 83% accuracy obtained by multi-network analysis under assumptions of stationary FC. This suggests that while many ICN-window size pairings may not exhibit strong disease discriminability individually, as evidenced in Figure 2 (right), there may exist important combinations of ICNs at certain time scales that can well characterize disease. Effectively, learning these modulation patterns allows for a whole-brain dynamic model of disease, where both stationary connectivity differences and associated functional compensation are captured. In other words, multiple-network dynamic FC approaches may be able to simultaneously describe both high-level aberrant connectivity and the resulting functional modulation of other brain networks in response. It will be interesting to see what multiple-network models come about in future analyses of the functional dynamics of disease.

4 Conclusion

In this study, we combined dynamic functional connectivity features from multiple networks to enhance the diagnosis of childhood autism. By using FC features over a wide range of time scales, our approach was able to substantially increase ASD classification when compared to using static FC features. Likewise, we showed that integrated network classification using a multiple kernel SVM approach has higher diagnostic potential when dynamic connectivity is considered. From our results, we conclude that incorporating different time scales for different ICNs into multi-network FC analysis provides an important, only now explored, perspective in our understanding of different disease mechanisms.

References

1. Beckmann, C., Mackay, C., Filippini, N.: SM, S.: Group comparison of resting-state fmri data using multi-subject ica and dual regression. In: OBHM (2009)
2. Chang, C., Glover, G.H.: Time–frequency dynamics of resting-state brain connectivity measured with fmri. NeuroImage 50(1), 81–98 (2010)

3. Chao-Gan, Y., Yu-Feng, Z.: Dparsf: a matlab toolbox for pipeline data analysis of resting-state fmri. Front. Sys. Neurosci. 4 (2010)
4. Di Martino, A., Yan, C., Li, Q., Denio, E., Castellanos, F., Alaerts, K., Anderson, J., Assaf, M., Bookheimer, S., Dapretto, M., et al.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatr. (2013)
5. Elton, A., Alcauter, S., Gao, W.: Network connectivity abnormality profile supports a categorical-dimensional hybrid model of adhd. Human Brain Mapping, n/a–n/a (2014)
6. Fox, M.D., Raichle, M.E.: Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nat. Rev. Neurosci. 8(9), 700–711 (2007)
7. Garrity, A., Pearlson, G., McKiernan, K., Lloyd, D., Kiehl, K., Calhoun, V.: Aberrant default mode functional connectivity in schizophrenia. Am. J. Psychiat. 164(3), 450–457 (2007)
8. Greicius, M.D., Srivastava, G., Reiss, A.L., Menon, V.: Default-mode network activity distinguishes alzheimer's disease from healthy aging: evidence from functional mri. P. Natl. Acad. Sci. USA 101(13), 4637–4642 (2004)
9. Hutchison, R.M., Womelsdorf, T., Allen, E.A., Bandettini, P.A., Calhoun, V.D., Corbetta, M., Penna, S.D., Duyn, J., Glover, G., Gonzalez-Castillo, J., et al.: Dynamic functional connectivity: Promises, issues, and interpretations. NeuroImage (2013)
10. Kelly Jr., R.E., Alexopoulos, G.S., Wang, Z., Gunning, F.M., Murphy, C.F., Morimoto, S.S., Kanellopoulos, D., Jia, Z., Lim, K.O., Hoptman, M.J.: Visual inspection of independent components: defining a procedure for artifact removal from fmri data. Journal of Neuroscience Methods 189(2), 233–245 (2010)
11. Kennedy, D.P., Adolphs, R.: The social brain in psychiatric and neurological disorders. Trends. Cogn. Sci. 16(11), 559–572 (2012)
12. Koshino, H., Carpenter, P.A., Minshew, N.J., Cherkassky, V.L., Keller, T.A., Just, M.A.: Functional connectivity in an fmri working memory task in high-functioning autism. NeuroImage 24(3), 810–821 (2005)
13. Leonardi, N., Richiardi, J., Gschwind, M., Simioni, S., Annoni, J.M., Schluep, M., Vuilleumier, P., Van De Ville, D.: Principal components of functional connectivity: A new approach to study dynamic brain connectivity during rest. NeuroImage 83, 937–950 (2013)
14. Ma, S., Calhoun, V.D., Phlypo, R., Adalı, T.: Dynamic changes of spatial functional network connectivity in healthy individuals and schizophrenia patients using independent vector analysis. NeuroImage (2014)
15. Mann, H.B., Whitney, D.R., et al.: On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. 18(1), 50–60 (1947)
16. Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E.: Spurious but systematic correlations in functional connectivity mri networks arise from subject motion. NeuroImage 59(3), 2142–2154 (2012)
17. Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., et al.: Advances in functional and structural mr image analysis and implementation as fsl. NeuroImage 23, S208–S219 (2004)
18. Sridharan, D., Levitin, D.J., Menon, V.: A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. P. Natl. A. Sci. 105(34), 12569–12574 (2008)
19. Uddin, L.Q., Supekar, K., Lynch, C.J., Khouzam, A., Phillips, J., Feinstein, C., Ryali, S., Menon, V.: Salience network–based classification and prediction of symptom severity in children with autism. JAMA Psychiatry 70(8), 869–879 (2013)
20. Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D.: Multimodal classification of alzheimer's disease and mild cognitive impairment. NeuroImage 55(3), 856–867 (2011)

Deriving a Multi-subject Functional-Connectivity Atlas to Inform Connectome Estimation

Ronald Phlypo[1,2], Bertrand Thirion[1,2], and Gaël Varoquaux[1,2]

[1] Parietal Team, Inria Saclay-Île-de-France, Saclay, France
{ronald.phlypo,gael.varoquaux,bertrand.thirion}@inria.fr
[2] CEA, DSV, I^{2}BM, Neurospin bât 145, 91191 Gif-Sur-Yvette, France

Abstract. The estimation of functional connectivity structure from functional neuroimaging data is an important step toward understanding the mechanisms of various brain diseases and building relevant biomarkers. Yet, such inferences have to deal with the low signal-to-noise ratio and the paucity of the data. With at our disposal a steadily growing volume of publicly available neuroimaging data, it is however possible to improve the estimation procedures involved in connectome mapping. In this work, we propose a novel learning scheme for functional connectivity based on sparse Gaussian graphical models that aims at minimizing the bias induced by the regularization used in the estimation, by carefully separating the estimation of the model support from the coefficients. Moreover, our strategy makes it possible to include new data with a limited computational cost. We illustrate the physiological relevance of the learned prior, that can be identified as a functional connectivity atlas, based on an experiment on 46 subjects of the Human Connectome Dataset.

Keywords: functional connectivity, sparse Gaussian graphical models.

1 Functional Connectivity and Rest fMRI

Functional connectivity (FC) is a simple measure of the interactions between brain regions. First introduced for electro-physiological recordings, such as spike-train recordings or electro-encephalography, it is now commonly used in spatially-resolved neuro-imaging modalities despite their poor temporal resolution at the scale of a second or more. Reported first as an alternative tool for positron emission tomography analysis [1], it has become a prominent tool in functional magnetic resonance imaging (fMRI) analysis. Functional connectivity has initiated a paradigm shift in fMRI studies, since it makes possible to scan subjects without engaging them in a controlled task. This is especially important when dealing with non-cooperative subjects, such as children or patients with neuropathologic disorders [2, among others]. In addition, it creates an opportunity to study the brain as a whole, through its global interactions, rather than through local effects.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 185–192, 2014.

Although FC is an established framework in neuroimaging, there is still no consensus on how to measure it from the data [3]. The connectivity pattern can be seen as a graph with its nodes (vertices) and connections (edges). Nodes are associated with elementary brain regions, between which connection strengths are then estimated, and the graph structure as a whole is referred to as a functional connectome. Nodes may either be predefined regions of interest (ROIs) or voxels [4], or may consist of distributed structures estimated from the data [5,6]. Once the nodes have been defined, edge strengths are inferred either through Pearson's correlation coefficients [4], partial correlation coefficients [7], or mutual information [8]. Characteristic network properties are then derived from these graph structures based on graph theoretic measures such as small-worldness, modularity, *etc* [9].

The central challenge in functional connectome estimation is that it is ill-posed, due to the small number of temporal samples in typical datasets. Regularization, *e.g.* with a sparsity assumption in the connections, is critical to recover connections from the data [3], yet at the expense of a bias on the estimated connectome. Ng *et al* [10] have shown that introducing a non-uniform connectivity prior derived from anatomical connectivity greatly improves estimation. However, the matching between functional and anatomical connectivity is often very imperfect [11]. For this reason, it is important to construct connectivity priors from functional connectivity information itself.

Here, we use a group of subjects to build such a prior in the form of an atlas that characterizes the stability of brain connections. Our estimation method is based on the precision matrix (inverse of correlation matrix, related to the partial correlation coefficients), since it has proven sensitive to correctly detect existing connections in a graph structure [3]. In contrast to existing stability studies, we do not focus on global properties of the graph, as in [9,12]. Rather, we aim at quantifying the probability of a connection being present between two pre-defined regions by means of a group study, and use this as a prior in the estimation of single subject FC. To alleviate the bias of sparsity-inducing penalization, we use a two-stage estimation procedure to obtain edge strengths.

Here, we focus on the qualitative properties of the model, which sketches a new generation of FC atlases, and leave quantitative validation for future work.

2 Data Used in the Study

In this contribution we use resting state data as provided by the minimal preprocessed data [13] of the human connectome project (HCP) [14] (release Q2). This comprises 46 subjects. All data were recorded in 3T MRI scanners with a repetition time of 720 ms. Data have been appropriately masked and motion corrected using the estimated motion parameters provided by HCP. In addition, we corrected for confounds using high variance confound regression [15] based on five principal components of 2 % highest variance voxel time series of whole brain image, and using five principal components of ventricle and white matter voxels (using standard masks provided in FSL, the mask of the latter eroded by 1 voxel to avoid overlap with gray matter).

3 Learning the Functional Connectivity Model

3.1 Graph Representation of Functional Connectivity

Given a set of p brain regions, the goal is to estimate an undirected graphical model $\mathcal{G}(V, E, \Omega)$ as a connectome representation. The vertex set $V = \{v_i, i \in [[p]]\}$ represents weighted averages of time series with respect to each of the p given ROIs (here a smooth positive function over a bounded subset of $\mathbb{R}^3$). Weights reflect '*probabilities*' of a voxel to belong to the given ROI. In this work, we have chosen the Harvard-Oxford lateralised probabilistic cortical atlas ($p = 96$) [16] (see, e.g., [12] for ROI abbreviations and acronyms used in this contribution).

The edge set $E = \{e_{ij}, (i,j) \in [[p]]^2\}$ contains e_{ij} if and only if the ROIs represented by v_i and v_j are functionally connected, given all other $v_k, k \notin \{i,j\}$. Edge weights $\Omega = \{\omega(e_{ij}), e_{ij} \in E\}$ represent signed connectivity strengths. Negative strengths correspond to a phase inversion between the time series of interest. When modeling FC, it is often assumed that the graph is sparse, which corresponds to a cardinality of E being (largely) inferior to $p(p-1)/2$. This hypothesis is justified based on evidence provided by structural connectivity measurements [17].

3.2 Estimation of the Graph Model Parameters

Estimating a sparse graphical-model is an NP-hard problem, and we choose graphical lasso [18,19] as an appropriate convex relaxation of our problem. We represent the graphical model through its adjacency matrix $\mathbf{\Theta} \overset{\text{def.}}{=} (\theta_{ij})_{i,j}$, which under a Gaussian hypothesis corresponds to the rescaled precision matrix (inverse of the covariance matrix). Hence, $\theta_{ij} = \theta_{ji} = 0$ if and only if the two regions represented by v_i and v_j are independent, conditional on all other regions [20]. Graphical lasso combines the maximum likelihood estimate with a convex sparsity inducing penalty with an objective function given by[1]

$$\widehat{\mathbf{\Theta}}_E = \arg\min_{\mathbf{\Theta} \succ \mathbf{0}} \varphi_E(\mathbf{\Theta}) \overset{\text{def.}}{=} -\log\det(\mathbf{\Theta}) + \operatorname{trace}(\mathbf{\Theta}\mathbf{S}) + \|\mathbf{\Lambda} \circ \mathbf{\Theta}\|_1 \qquad (1)$$

where $\mathbf{S}$ is the sample correlation matrix of the time-series represented by the nodes in V. We minimize $\varphi_E(\mathbf{\Theta})$ of Eq. (1) over the space of symmetric positive definite matrices using the algorithm of [21], which guarantees $\widehat{\mathbf{\Theta}}_E \succ \mathbf{0}$ at any step of the algorithm, as such providing a procedure more robust against ill-conditioning of $\mathbf{S}$ than those originally proposed in [18,19].

The estimate $\widehat{\mathbf{\Theta}}_E$ is a biased estimate of $\mathbf{\Theta}$, and we only use it to infer the edge set E. We thus break our estimation scheme down into three steps: (i) estimation

[1] The original implementation of the penalization in graph lasso reads $\lambda\|\mathbf{\Theta}\|_1$, but since no penalization of the diagonal terms is required, we may write $\lambda\left\|\left(\mathbb{1}\mathbb{1}^\mathsf{T} - \mathrm{Id}_p\right) \circ \mathbf{\Theta}\right\|_1$, where $\mathbb{1}$ is a vector of ones in $\mathbb{R}^p$ and $\cdot \circ \cdot$ denotes the element-wise or Hadamard product. We use here an extension of this model, replacing $\lambda\left(\mathbb{1}\mathbb{1}^\mathsf{T} - \mathrm{Id}_p\right)$ by a symmetric matrix $\mathbf{\Lambda} = (\lambda_{ij})_{i,j}$, with $0 \leq \lambda_{ij} \leq 1$.

of the hyperparameter $\mathbf{\Lambda}$, (ii) estimation of the edge set E, and (iii) estimation of the edge weights or signed connectivity strengths $\mathbf{\Omega}$.
Step (i) will be discussed in the next section. Assuming for the moment that $\mathbf{\Lambda}$ is known, the last two steps are simply:
(ii) The edges e_{ij} in $\widehat{E}$ correspond to those entries of $\widehat{\mathbf{\Theta}}_E$ for which $\left|\widehat{\theta}_{ij}\right| > 0$.
(iii) Given the estimated edge set, we estimate edge strengths $\mathbf{\Omega} = (\omega_{ij})_{i,j}$ as the maximum likelihood estimates restricted to the edge set $\widehat{E}$ of the graph $\widehat{\mathcal{G}}$, based on the sample correlation matrix $\mathbf{S}$. Our estimator reads

$$\widehat{\mathbf{\Theta}}_\Omega = \arg\min_{\mathbf{\Theta}|_{\widehat{\mathcal{G}}} \succ \mathbf{0}} \varphi_\Omega(\mathbf{\Theta}) \stackrel{\text{def.}}{=} -\log\det(\mathbf{\Theta}) + \operatorname{trace}(\mathbf{\Theta}\mathbf{S}) \tag{2}$$

which is over the space of positive-definite, symmetric matrices with a given support set $\widehat{E}$. The restriction to the graph prohibits the use of straightforward matrix inversion, and instead we use iterative proportional scaling (IPS), iteratively solving plain maximum likelihood models alternated over maximal cliques of the graph $\widehat{\mathcal{G}}$ [20]. The obtained solution is an unbiased maximum likelihood estimator conditional on the support set of the graph, hence the only bias in our final model is the pruning of the edges. Given new data (a subject, session, or condition), we can infer the corresponding functional connectome.

3.3 Choosing the Prior Atlas

The choice of the regularizer $\mathbf{\Lambda}$ is essential to optimize the bias/variance compromise. A first approach consists in choosing $\mathbf{\Lambda} = \lambda\left(\mathbb{1}\mathbb{1}^\mathsf{T} - \mathrm{Id}_p\right)$. As in [21], this makes the problem low-dimensional. Hence, we use a repeated random jackknifing cross-validation scheme. As a score function, we opt for the negative log-likelihood, which is our main objective function. However, the extrapolation of that single parameter λ is an issue, since a given λ value corresponds to different cardinality of $\widehat{E}$ for different datasets (even if ratios such as $\lambda/\lambda_{\max}$ are considered, where $\lambda_{\max}$ is the critical regularizer that yields a trivial solution). Instead we propose to accumulate the support estimates across subjects and sessions. The intuition behind this approach is that noise in the data can cause instabilities in the graph lasso estimate. However, these instabilities are not reproducible across subjects, and the population distribution of the presence of edges is reliable. This approach is related to the "stability selection" strategy, also developed for sparse Gaussian graphical models [22]. Consider q subjects, each subject having 2 separate scan sessions [2].

Estimation of $\lambda^{(i)}$s. For each subject i, we use a grid search to estimate $\lambda^{(i)}$ as that λ in $\mathbf{\Lambda} = \lambda\left(\mathbb{1}\mathbb{1}^\mathsf{T} - \mathrm{Id}_p\right)$ for which $\widehat{\mathbf{\Theta}}_\Omega^{(i)}$ maximizes the likelihood generalization. The generalization capability of the likelihood can be approximated

[2] We consider the back-to-back recorded left-right and right-left phase encoding as a single session. To marginalize the effect of the phase encoding, we always sample both encodings evenly. Thus, if we state that n samples are jackknifed from a session, we actually mean that $n/2$ samples are jackknifed from each phase encoding associated with that session.

by using n samples from one session as a train session to estimate $\boldsymbol{\Theta}^{(i)}_{E^{(i)}}$; $E^{(i)}$ is then taken as the support in Eq. (2) to estimate $\boldsymbol{\Theta}^{(i)}_{\Omega^{(i)}}$ from n samples. We then use the n' complementary samples in the latter session to compute the likelihood of $\widehat{\boldsymbol{\Theta}}^{(i)}_{\Omega^{(i)}}$. This is repeated k times, and we report $\lambda^{(i)}$, as that λ that maximizes the average likelihood. In this study, we have used $n = 200$, $n' = 600$, $k = 5$, and for each repetition the session used for training has been chosen randomly.

Estimation of $\boldsymbol{\Lambda}$. Using the optimal $\lambda^{(i)}$ for subject i, we re-estimate $E^{(i)}$ from $n'' = 400$ samples, using Eq. (1) for each session, yielding $\widehat{E}^{(i,1)}$ and $\widehat{E}^{(i,2)}$. Suppose that all subjects (and sessions) are drawn from a single population with $\mathbb{P}_{ij}$ the marginal probability of observing edge e_{ij}. We simply estimate these probabilities by using maximum likelihood under marginal Bernoulli distributions on the edges, based on the $\{E^{(i,1)}, E^{(i,2)}\}$ estimates. We then define the penalization parameter $\boldsymbol{\Lambda}$ as $\mathbb{1}\mathbb{1}^\intercal - \mathbf{P}$, where $\mathbf{P} \overset{\text{def.}}{=} (\mathbb{P}_{ij})_{i,j}$. Note that, by construction, $\mathbf{P}$ is a symmetric matrix with ones on its diagonal.

$\boldsymbol{\Lambda}$ can then be used on a new subject as a penalty in the penalized maximum likelihood estimation Eq. (1). Importantly, it appears as a penalty, not a constraint, and it is thus a "soft" prior. In other words, it does not force the existence or absence of a connections; it simply promotes these choices.

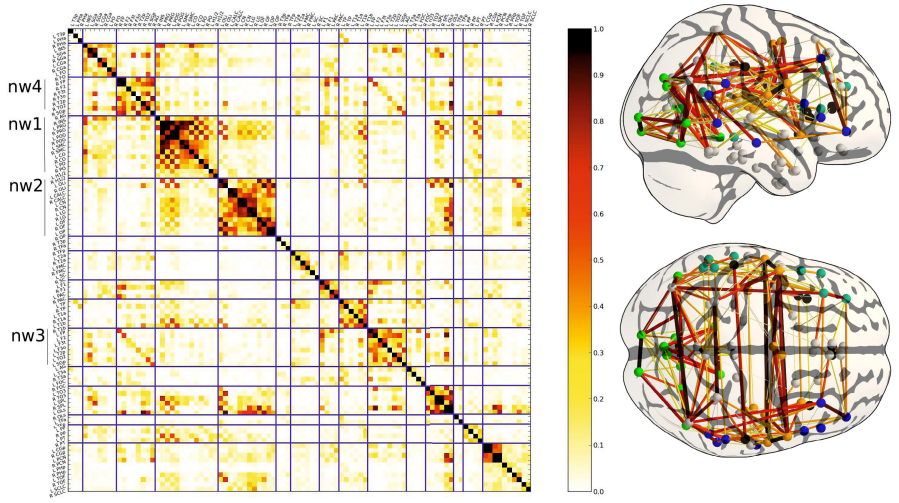


Fig. 1. Prior on the functional connectivity, noted $\mathbf{P}$ in the text: the coefficients in the matrix represent the frequencies of edge detections. ROI labels and their abbreviations, as well as a visualization of all networks can be found in an addendum to a preprint of this manuscript at `http://hal.inria.fr/hal-00991124`.

4 Resulting Atlas and Discussion

The probability estimates in $\mathbf{P}$ give a global view of the connectome in our training set of 46 subjects. It is important to note that the structure of $\mathbf{P}$ reflects the sparsity assumptions, which means that some connections have indeed never been 'observed' over the entire set of subjects, see Fig. 1. This connectivity prior matrix is very structured; it displays a checkerboard pattern: regions are most likely to be connected to a region in the same hemisphere than the corresponding one in the opposite hemisphere. In addition, visualized as a brain graph, it features well-known characteristics of brain connectivity, such as strong homologous inter-hemispheric or fronto-parietal connections.

To visualize the latent network structure, we re-organized the matrix $\mathbf{P}$ using hierarchical clustering with complete linkage, choosing the number of clusters that minimizes the silhouette criterion [23]. Clustering on probabilities of edges in the support of a graph gives FC networks (see Fig. 2) similar to those found in group analyses based on independent component maps, see, e.g., [5]. This suggests that the FC atlas has indeed captured valuable information from our reference population.

Other than having an interpretable graph representation, it should be emphasized that our approach provides a readily-usable subject-specific estimator, alleviating the need of a search over the parameter space (λ) for a new subject. As any medical-imaging atlases, our model incorporates accumulated knowledge. In contrast to group analyses that need a complete re-computation of the graph or the components upon incorporation of a new subject [6,24], we only need to update a single matrix, namely the estimator $\mathbf{P}$. Moreover, based on the observations of [10], we conjecture that the incorporation of the structured atlas-based prior will yield more accurate estimation of functional connectivity than traditional graph lasso estimators that enforce a uniform ℓ_1 shrinkage on all connections.

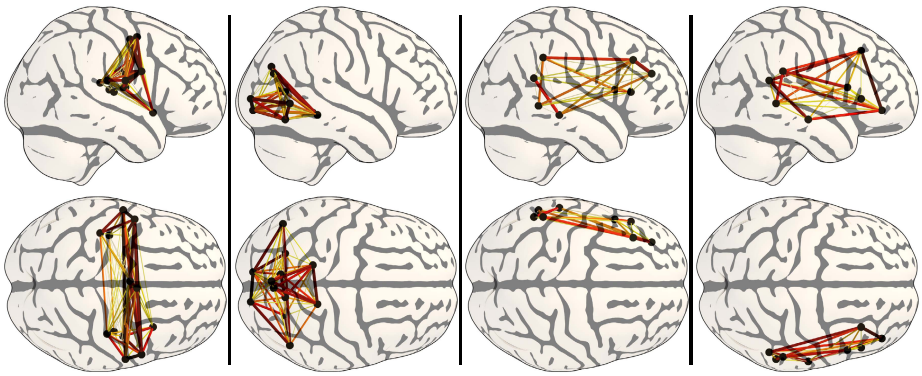


Fig. 2. Lateral and top views of the four networks containing the most nodes. From left to right we distinguish a motor network, a visual network, as well as two lateralised TPJ-to-IFG networks (TPJ: temporo-parietal junction – IFG: Inferior Frontal Gyrus).

5 Conclusions

Sparsifying penalties are the most promising tools to learning functional connectomes from noisy and scarce fMRI data. Injecting structure in this prior, favoring connections that are known to be likely, can greatly improve the ability of the estimation procedures to recover real connections. We have introduced a procedure to learn an atlas of brain functional connections by accumulating knowledge across subjects, associated with a subject-specific estimator of a sparse Gaussian graphical model that quantifies functional connectivity. Future work includes the validation of the atlas-based strategy for individual connectivity estimation and subjects connectivity comparison.

Acknowledgements. Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Uğurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. We acknowledge funding from the NiConnect project grant.

References

1. Friston, K.J., Frith, C.D., Liddle, P.F., Frackowiak, R.S.J.: Functional connectivity: the principal-component analysis of large (PET) data sets. Journal of Cerebral Blood Flow and Metabolism 13, 5–14 (1993)
2. Menon, V.: Developmental pathways to functional brain networks: emerging principles. Trends in Cognitive Sciences (2013)
3. Smith, S.M., Miller, K.L., Salimi-Khorshidi, G., Webster, M., Beckmann, C.F., Nichols, T.E., Ramsey, J.D., Woolrich, M.W.: Network modelling methods for fMRI. NeuroImage 54(2), 875–891 (2011)
4. Hayasaka, S., Laurienti, P.J.: Comparison of characteristics between region-and voxel-based network analyses in resting-state fMRI data. NeuroImage 50(2), 499–508 (2010)
5. Damoiseaux, J.S., Rombouts, S.A.R.B., Barkhof, F., Scheltens, P., Stam, C.J., Smith, S.M., Beckmann, C.F.: Consistent resting-state networks across healthy subjects. Proceedings of the National Academy of Sciences 103(37), 13848–13853 (2006)
6. Smith, S.M., Miller, K.L., Moeller, S., Xu, J., Auerbach, E.J., Woolrich, M.W., Beckmann, C.F., Jenkinson, M., Andersson, J., Glasser, M.F., Van Essen, D.C., Feinberg, D.A., Yacoub, E.S., Ugurbil, K.: Temporally-independent functional modes of spontaneous brain activity. Proceedings of the National Academy of Sciences 109(8), 3131–3136 (2012)
7. Marrelec, G., Krainik, A., Duffau, H., Pélégrini-Issac, M., Lehéricy, S., Doyon, J., Benali, H.: Partial correlation for functional brain interactivity investigation in functional MRI. NeuroImage 32(1), 228–237 (2006)
8. Ma, S., Calhoun, V.D., Phlypo, R., Adalı, T.: Dynamic changes of spatial functional network connectivity in healthy individuals and schizophrenia patients using independent vector analysis. NeuroImage 90, 196–206 (2014)

9. Rubinov, M., Sporns, O.: Complex network measures of brain connectivity: Uses and interpretations. NeuroImage 52(3), 1059–1069 (2010); Computational Models of the Brain
10. Ng, B., Varoquaux, G., Poline, J.B., Thirion, B.: A novel sparse graphical approach for multimodal brain connectivity inference. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part I. LNCS, vol. 7510, pp. 707–714. Springer, Heidelberg (2012)
11. Ng, B., Varoquaux, G., Poline, J.B., Thirion, B.: Implications of inconsistencies between fMRI and dMRI on multimodal connectivity estimation. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part III. LNCS, vol. 8151, pp. 652–659. Springer, Heidelberg (2013)
12. Wang, J.H., Zuo, X.N., Gohel, S., Milham, M.P., Biswal, B.B., He, Y.: Graph theoretical analysis of functional brain networks: Test-retest evaluation on short- and long-term resting-state functional MRI data. PLoS One 6(7), e21976 (2011)
13. Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., Van Essen, D.C., Jenkinson, M.: The minimal preprocessing pipelines for the human connectome project. NeuroImage 80, 105–124 (2013); Mapping the Connectome
14. Smith, S.M., Beckmann, C.F., Andersson, J., Auerbach, E.J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D.A., Griffanti, L., Harms, M.P., Kelly, M., Laumann, T., Miller, K.L., Moeller, S., Petersen, S., Power, J., Salimi-Khorshidi, G., Snyder, A.Z., Vu, A.T., Woolrich, M.W., Xu, J., Yacoub, E., Uğurbil, K., Van Essen, D.C., Glasser, M.F.: Resting-state fMRI in the human connectome project. NeuroImage 80, 144–168 (2013); Mapping the Connectome
15. Behzadi, Y., Restom, K., Liau, J., Liu, T.T.: A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. NeuroImage 37(1), 90–101 (2007)
16. Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J.: An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31(3), 968–980 (2006)
17. Deligianni, F., Varoquaux, G., Thirion, B., Sharp, D.J., Ledig, C., Leech, R., Rueckert, D.: A framework for inter-subject prediction of functional connectivity from structural networks. IEEE Transactions on Medical Imaging (August 2013)
18. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics 3, 432–441 (2008)
19. Banerjee, O., El Ghaoui, L., d'Aspremont, A.: Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. Journal of Machine Learning Research 9, 485–516 (2008)
20. Lauritzen, S.L.: Graphical Models. Oxford Statistical Science, vol. 17. Clarendon Press (1996)
21. Mazumder, R., Hastie, T.: The graphical lasso: new insights and alternatives. Electronic Journal of Statistics 6, 2125–2149 (2012)
22. Meinshausen, N., Bühlmann, P.: Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72(4), 417–473 (2010)
23. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65 (1987)
24. Ng, B., Varoquaux, G., Poline, J.B., Thirion, B.: A novel sparse group Gaussian graphical model for functional connectivity estimation. In: Information Processing in Medical Imaging, Asilomar, États-Unis (June 2013)

Discriminative Sparse Connectivity Patterns for Classification of fMRI Data

Harini Eavani[1], Theodore D. Satterthwaite[2], Raquel E. Gur[2], Ruben C. Gur[2], and Christos Davatzikos[1]

[1] Center for Biomedical Image Computing and Analytics, University of Pennsylvania
[2] Brain Behavior Laboratory, University of Pennsylvania

Abstract. Functional connectivity using resting-state fMRI has emerged as an important research tool for understanding normal brain function as well as changes occurring during brain development and in various brain disorders. Most prior work has examined changes in pair-wise functional connectivity values using a multi-variate classification approach, such as Support Vector Machines (SVM). While it is powerful, SVMs produce a dense set of high-dimensional weight vectors as output, which are difficult to interpret, and require additional post-processing to relate to known functional networks. In this paper, we propose a joint framework that combines network identification and classification, resulting in a set of networks, or Sparse Connectivity Patterns (SCPs) which are functionally interpretable as well as highly discriminative of the two groups. Applied to a study of normal development classifying children vs. adults, the proposed method provided accuracy of 76%(AUC= 0.85), comparable to SVM (79%,AUC=0.87), but with dramatically fewer number of features (50 features vs. 34716 for the SVM). More importantly, this leads to a tremendous improvement in neuro-scientific interpretability, which is specially advantageous in such a study where the group differences are wide-spread throughout the brain. Highest-ranked discriminative SCPs reflect increases in long-range connectivity in adults between the frontal areas and posterior cingulate regions. In contrast, connectivity between the bilateral parahippocampal gyri was decreased in adults compared to children.

1 Introduction

Functional connectivity, defined as the amount of correlation between observed BOLD time-series, has been widely applied to study the large-scale functional architecture of the human brain in both health, disease, and development. Many studies that examine changes in pair-wise connectivity use multi-variate methods, such as SVMs, with the vectorized correlation matrices as features [1]. Applied to classification, the l_2-regularized SVM results in a list of connections or edges that are discriminative of the two groups. However, high-dimensional patterns are very difficult to parse and interpret, requiring additional processing and

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 193–200, 2014.

analysis based on prior knowledge of known functional network labels [2]. To address this issue, feature selection methods [3] or l_1-regularized SVM can be used, which produces a sparse discriminative pattern. However, it is known to ignore features that are highly correlated (and therefore redundant for classification), which could be neurobiologically relevant.

Alternately, a complementary set of analysis that can be performed begins with network identification using independent component analysis (ICA). ICA commonly identifies motor, visual, sub-cortical and default mode networks. These networks can then be analyzed for changes in spatial extent or average connectivity between the two groups. However, in such a method, network identification is performed independently of classification. This can be problematic since not all networks are necessarily different in the two groups - requiring an exhaustive search in order to find group differences. Such a method seeks to represent functional activity in general, and does not aim specifically to find networks that relate to the classification task.

In order to address the above limitations, in this paper, we propose a joint framework that combines network identification with classification, resulting in functionally meaningful networks that are highly discriminative of the two groups. Our method is akin to previous work in structural MRI [4], presenting a joint generative-discriminative formulation. We propose the use of sparse decompositions for network identification, resulting in a small set of *Sparse Connectivity Patterns* (SCPs) that are highly discriminative of the two groups, as well as functionally interpretable. Thus, this method reduces the high dimensionality of the connectivity data to a small set of strongly correlated regions, that are relevant to the classification task. In addition, as opposed to vectorizing the correlation matrices [1] the proposed method exploits the *positive-semi-definite* (PSD) property of the correlation matrix [5], in order to find stable SCPs.

To test its performance, we use the proposed method to investigate functional connectivity differences between children and young adults. Section 2 provides details of the proposed method. Section 3 describes the results and discusses the merits and drawbacks of the proposed method compared to other methods. Conclusions and future work are provided in section 4.

2 Identification of Discriminative Sparse Connectivity Patterns

A schematic diagram illustrating our method is shown in Figure 1. Given P regions, the input to the method is size $P \times P$ correlation matrices $\mathbf{\Sigma}_n \succeq 0$, and the associated binary group membership y_n, for each subject n, $n = 1, 2, \ldots, N$. We would like to find smaller networks, or SCPs common to all the subjects, such that the total connectivity within each network contributes to the two-group classification. Our formulation jointly optimizes two objectives: (1) Identification of SCPs (2) Learning discriminative SCPs. In the following sub-sections we provide the details of each, followed by the joint optimization strategy.

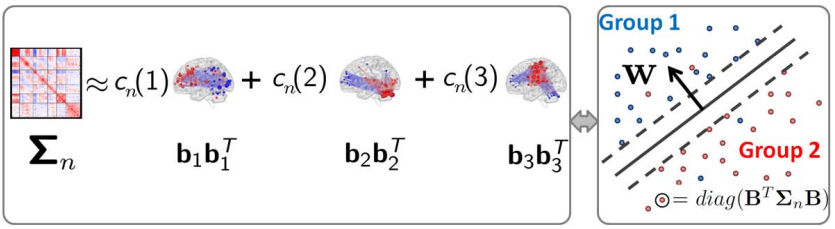


Fig. 1. Schematic illustrating the joint framework. Panel to the left describes the SCP identification term, which factorizes connectivity matrices $\boldsymbol{\Sigma}_n$ of each subject n into a set of common SCPs $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_K]$ and its associated coefficients. Panel to the right illustrates a linear SVM, which uses the total absolute connectivity values $diag(\mathbf{B}^T\boldsymbol{\Sigma}_n\mathbf{B})$ as input features to classify two groups, resulting in the hyperplane $\mathbf{w}$.

2.1 Identification of SCPs

We would like to find SCPs common to all the subjects, such that a non-negative combination of SCPs generates the correlation matrix $\boldsymbol{\Sigma}_n$, for each subject n. We represent each SCP by a vector of region-weights $\mathbf{b}_k$, where $-1 \preceq \mathbf{b}_k \preceq 1$, $\mathbf{b}_k \in \mathbf{R}^P$, reflecting the membership of the regions to SCP k. If two regions in $\mathbf{b}_k$ have the same sign, then they are positively correlated and opposing sign reflects anti-correlation. Thus, the rank-one matrix $\mathbf{b}_k\mathbf{b}_k^T$ reflects the correlation behavior of SCP k. To retain the P.S.D. nature of the correlation matrix, we would like to approximate each matrix $\{\boldsymbol{\Sigma}_n\}_{n=1}^N$ by a non-negative combination of these sub-networks $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_K]$. Thus, we want $\boldsymbol{\Sigma}_n \approx \sum_{k=1}^K c_n(k)\mathbf{b}_k\mathbf{b}_k^T = \mathbf{B}\ \mathrm{diag}(\mathbf{c}_n)\ \mathbf{B}^T$, where $\mathrm{diag}(\mathbf{c}_n)$ denotes a diagonal matrix with values $\mathbf{c}_n \in \mathbf{R}_+^K$ along the diagonal.

We quantify the approximation above using the frobenius norm. Then the loss function $\mathcal{G}$ takes the form:

$$\mathcal{G}(\mathbf{B}, \mathbf{C}) = \sum_{n=1}^{N} \left|\left|\boldsymbol{\Sigma}_n - \mathbf{B}\ \mathrm{diag}(\mathbf{c}_n)\ \mathbf{B}^T\right|\right|_F^2 \quad (1)$$

In general, a *blind-decomposition* problem such as the above is ill-posed, i.e., multiple optima exist, and the solutions are not stable with respect to the noise in the data. To make the results stable, additional constraints need to be imposed on the matrix factors. Since known functional networks such as the visual or motor networks have small spatial extent compared to the whole brain, we use spatial sparsity as a constraint, similar to our prior work [6]. Hence we impose spatial sparsity on the SCPs by restricting the l_1-norm of $\mathbf{b}_k$ to less than a constant value λ.

2.2 Learning Discriminative SCPs

Our objective is to find SCPs that can reconstruct that data as well as act as a discriminative basis that can classify the two groups. In other words, each SCP

consists of regions which have the following properties (1) they are strongly positively or negatively correlated (2) the total absolute correlation between all the regions within an SCP contributes towards the group difference. The first property is modeled in the previous section; in this section we describe the discriminative term that models the second property.

Given an SCP $\mathbf{b}_k$, the scalar value $\mathbf{b}_k^T \mathbf{\Sigma}_n \mathbf{b}_k$ measures the total absolute correlation between all the regions within the SCP for a given subject n. Computed for all SCPs, the K-dimensional vector $diag(\mathbf{B}^T \mathbf{\Sigma}_n \mathbf{B})$ serves as the subject-specific measure that can be used in a multi-variate SVM framework. We use the squared hinge loss and l_2 regularization for the K-dimensional SVM hyperplane $\mathbf{w}$. The cost function for the discriminative term is:

$$\mathcal{D}(\mathbf{B}, \mathbf{w}) \ = \sum_{n=1}^{N} \left(1 - y_n \mathbf{w}^T \ diag\left(\mathbf{B}^T \mathbf{\Sigma}_n \mathbf{B}\right)\right)_+^2 \tag{2}$$

where y_n is the binary group label for subject n, and the subscript + denotes the positive part of the argument.

2.3 Joint Optimization Framework

Bringing the terms in Eqns. 1 and 2 together, along with the l_2 regularizer for $\mathbf{w}$, we have the optimization problem

$$\begin{aligned} &\underset{\mathbf{B},\mathbf{C},\mathbf{w}}{\text{minimize}} \ (1-\mu)\mathcal{G}(\mathbf{B},\mathbf{C}) + \mu\mathcal{D}(\mathbf{B},\mathbf{w}) + ||\mathbf{w}||_2^2 \\ &\text{subject to} \\ &\qquad -1 \le \mathbf{b}_k \le 1, \quad ||\mathbf{b}_k||_1 \le \lambda, \quad k = 1,\ldots,K \\ &\qquad\quad \mathbf{c}_n \ge 0, \qquad\quad n = 1,\ldots,N \end{aligned} \tag{3}$$

where μ is the relative weighting fraction between the two terms. A value of $\mu = 0$ produces purely reconstructive SCPs. We use alternating minimization to iteratively solve for $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{w}$. We use a projected gradient method [4] for $\mathbf{B}$ and $\mathbf{C}$ and the *libSVM* solver for $\mathbf{w}$ [7]. The parameter μ, which controls the trade-off between the two terms, will be linearly increased from a value of 0 to 1 during the iterative process. This ensures that the SCPs generated during the first few iterations are mainly reconstructive, which tend to be more stable.

Model Parameters The free parameters of the proposed method are the number of SCPs K, and the sparsity level λ. Using grid search, for every pair of values in $K \in \{10, 20, \ldots\}$ and $\lambda \in \{0.01, 0.02, \ldots, 0.1\} * P$, we will use repeated five-fold cross-validation to find the optimal set of parameters.

3 Application to Study of Development

3.1 Data and Pre-processing

Data used here was drawn from the PNC database [8]. Our data consists of 91 children (age = 10.38 ± 1.01 yrs.) and 84 young adults (age= 20.21 ± 0.84 yrs.).

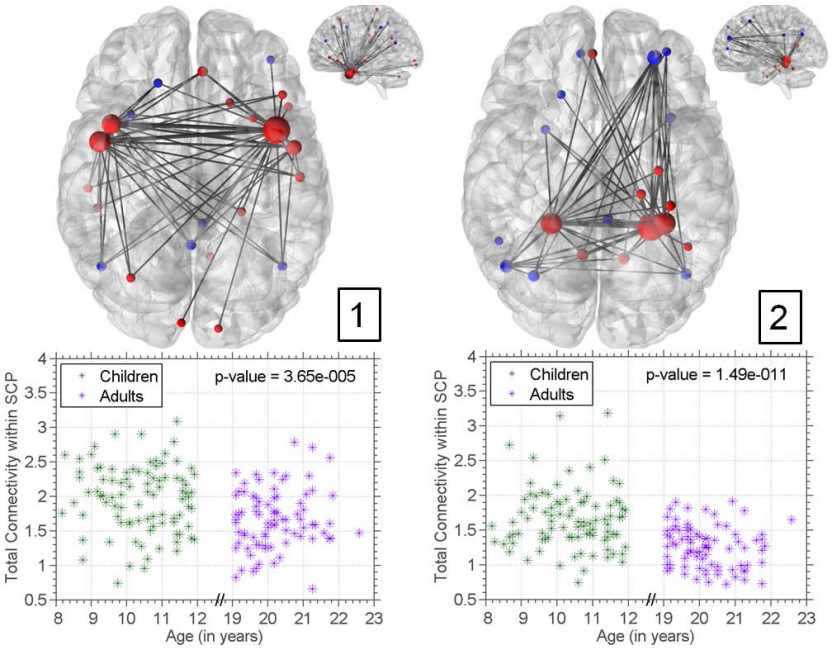


Fig. 2. Top two SCPs whose total connectivity is stronger in children. 3D brain rendering (top) displays location of regions belonging to the SCP. Opposing colors (red/blue) indicate anti-correlation between the regions. Size of nodes reflect SCP values $\mathbf{b}_k$. Corresponding graphs (bottom) plot total connectivity within SCP for each subject vs. subject age. Uni-variate p-value scores comparing total connectivity between two groups are also shown.

As head motion is a known confound that correlates with age effects, we matched the two groups on motion, measured using the mean relative displacement [1,9] ($p = 0.79$). As described elsewhere in detail [8], Blood Oxygen Level Dependent (BOLD) fMRI was acquired using a whole-brain, echoplanar (EPI) sequence with the following parameters: 124 volumes, TR 3000 ms, TE 32 ms, effective voxel resolution 3.0x3.0x3.0mm.

Subject-level BOLD images fMRI images were affinely registered to the T1 image, followed by non-linear registration to the MNI 152 template. We used the 264 nodes defined in [10] for our experiments. Time-series data was preprocessed using a validated confound regression procedure [11]. Averaged time-courses corresponding to these 264 nodes were used to compute a symmetric Pearson correlation matrix $\mathbf{\Sigma}_n \in \mathcal{S}_+^P$ for each subject.

3.2 Results Using Proposed Method

The results of the cross-validation provided an operating point of $K = 50, \lambda = 0.03P$ (roughly 10 nodes per SCP) at which the results are generalizable within

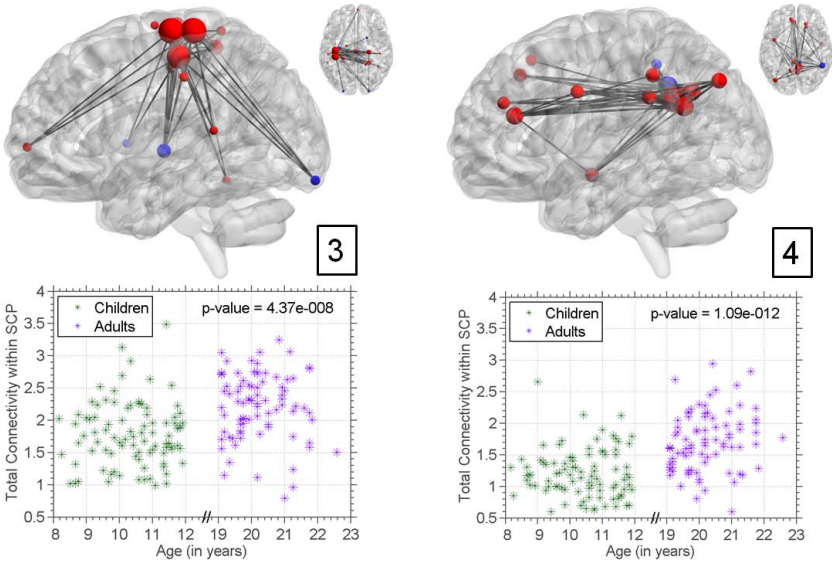


Fig. 3. Top two discriminative SCPs whose total connectivity is stronger in adults

sub-samples of the dataset. The cross-validation accuracy saturates at higher values of K. For these values, the proposed method gave an average classification accuracy of $76.3 \pm 7.08\%$ between children vs. young adults. The most discriminative SCPs are shown in Figures 2 and 3, rendered using *BrainNet* Viewer [12].

Children demonstrated stronger connectivity within the two SCPs, displayed in Figure 2. Positive correlations between the bilateral temporal poles and the parahippocampal gyri are significantly stronger in children, than in adults. This pattern is consistent with an increased degree of lateralization of temporal connectivity with development [13]. In contrast, adults demonstrated stronger connectivity within two SCPs including the motor network and the default mode network (Figure 3). These patterns are consistent with patterns of network segregation, whereby connectivity within certain major networks including the motor network and the default mode network increase with development [2,1]. Of the four SCPs, strongest p-value difference is exhibited by SCP 4, which shows increased anterior-posterior connectivity between two default-mode regions - medial Pre-Frontal Cortex (mPFC) and the Posterior Cingulate Cortex (PCC). This finding has been consistently found in other studies [2], further adding to the validity of our method.

3.3 Comparison with Other Methods

We compared our method with four alternate approaches: (1) l_2-regularized l_2-loss linear SVM [7] with pair-wise correlation values (2) l_1-regularized l_2-loss

Table 1. Accuracy and AUC values for all methods

Method	Acc (%)	AUC	# features
Proposed	76.3 ± 7.08	0.85 ± 0.06	50
l_2-SVM	79 ± 1.45	0.87 ± 0.01	34716
l_1-SVM	74 ± 2.25	0.81 ± 0.01	139
PCA+SVM	65.2 ± 2.30	0.67 ± 0.01	180
ICA+SVM	64.8 ± 1.42	0.71 ± 0.03	210

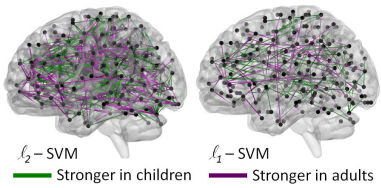


Fig. 4. SVM weight vector for l_2-SVM (left) and l_1-SVM (right)

linear SVM (3) Principal Component Analysis (PCA), followed by classification (4) Independent Component Analysis (ICA), followed by classification. The first and second methods are purely discriminative, as they do not perform network identification. The third and fourth methods use unsupervised network identification methods, followed by classification using the total absolute connectivity values as features. The number of components K (in PCA and ICA) and the cost parameter for the SVM was chosen using cross-validation.

The classification performance for all the methods is reported in Table 1. The un-supervised PCA and ICA methods perform poorly. Of the three methods, l_2-SVM provides a slightly better performance compared to the proposed method, although the difference in accuracies between the two methods is insignificant ($p = 0.0625$). The marginally higher accuracy provided by the SVM is due to the 1000-fold increase in the number of features used, leading to a complete loss of interpretability. This is illustrated in Figure 4, which displays the l_2-SVM weight vector. The weight vector for l_1-SVM is also shown in the same figure. While the l_1 penalty does dramatically reduce the number of features used, it does not necessarily alleviate the issue of non-interpretability. As explained earlier, strongly correlated features (connections) that are redundant to the classification are dropped. In contrast, the generative term $\mathcal{G}$ within the proposed method tends to retain these features by allocating them to the same SCP. Thus, a whole-brain discriminative pattern is split into multiple SCPs based on the dependencies between the connection strengths, allowing results to be interpreted within the context of known functional brain networks.

4 Conclusion

In this paper we presented a framework that performs supervised dimensionality reduction for functional connectivity data such that the discriminative pattern between two groups is preserved. Results demonstrate improved neurobiological interpretablity compared to purely discriminative approaches without a significant loss of accuracy. SCPs that discriminate children from young adults are consistent with previously-described reports of increased temporal lateralization and functional network segregation. Future work will extend this method to continuous labels within a regression framework.

References

1. Satterthwaite, T.D., Wolf, D.H., Ruparel, K., Erus, G., Elliott, M.A., Eickhoff, S.B., Gennatas, E.D., Jackson, C., Prabhakaran, K., Smith, A., et al.: Heterogeneous impact of motion on fundamental patterns of developmental changes in functional connectivity during youth. NeuroImage 83, 45–57 (2013)
2. Fair, D.A., Cohen, A.L., Power, J.D., Dosenbach, N.U., Church, J.A., Miezin, F.M., Schlaggar, B.L., Petersen, S.E.: Functional brain networks develop from a local to distributed organization. PLoS Computational Biology 5(5), e1000381 (2009)
3. Venkataraman, A., Kubicki, M., Westin, C.F., Golland, P.: Robust feature selection in resting-state fmri connectivity based on population studies. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 63–70. IEEE (2010)
4. Batmanghelich, N., Taskar, B., Davatzikos, C.: Generative-discriminative basis learning for medical imaging. IEEE Transactions on Medical Imaging 31(1), 51–69 (2012)
5. Varoquaux, G., Baronnet, F., Kleinschmidt, A., Fillard, P., Thirion, B.: Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part I. LNCS, vol. 6361, pp. 200–208. Springer, Heidelberg (2010)
6. Eavani, H., Satterthwaite, T.D., Gur, R.E., Gur, R.C., Davatzikos, C.: Unsupervised learning of functional network dynamics in resting state fmri. In: Gee, J.C., Joshi, S., Pohl, K.M., Wells, W.M., Zöllei, L. (eds.) IPMI 2013. LNCS, vol. 7917, pp. 426–437. Springer, Heidelberg (2013)
7. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011)
8. Satterthwaite, T.D., Elliott, M.A., Ruparel, K., Loughead, J., Prabhakaran, K., Calkins, M.E., Hopson, R., Jackson, C., Keefe, J., Riley, M., et al.: Neuroimaging of the philadelphia neurodevelopmental cohort. NeuroImage 86, 544–553 (2014)
9. Van Dijk, K.R., Sabuncu, M.R., Buckner, R.L.: The influence of head motion on intrinsic functional connectivity mri. Neuroimage 59(1), 431–438 (2012)
10. Power, J.D., Cohen, A.L., Nelson, S.M., Wig, G.S., Barnes, K.A., Church, J.A., Vogel, A.C., Laumann, T.O., Miezin, F.M., Schlaggar, B.L., Petersen, S.E.: Functional network organization of the human brain. Neuron 72(4), 665–678 (2011)
11. Satterthwaite, T.D., Elliott, M.A., Gerraty, R.T., Ruparel, K., Loughead, J., Calkins, M.E., Eickhoff, S.B., Hakonarson, H., Gur, R.C., Gur, R.E., et al.: An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. Neuroimage 64, 240–256 (2013)
12. Xia, M., Wang, J., He, Y.: Brainnet viewer: A network visualization tool for human brain connectomics. PloS One 8(7), e68910 (2013)
13. Zuo, X.N., Kelly, C., Di Martino, A., Mennes, M., Margulies, D.S., Bangaru, S., Grzadzinski, R., Evans, A.C., Zang, Y.F., Castellanos, F.X., et al.: Growing together and growing apart: regional and sex differences in the lifespan developmental trajectories of functional homotopy. The Journal of Neuroscience 30(45), 15034–15043 (2010)

MesoFT: Unifying Diffusion Modelling and Fiber Tracking

Marco Reisert[1], V.G. Kiselev[1], Bibek Dihtal[1], Elias Kellner[1], and D.S. Novikov[2]

[1] Department of Diagnostic Radiology, Medical Physics, University Medical Center Freiburg, Breisacher Street 60a, 79106 Freiburg, Germany
[2] Bernard and Irene Schwartz Center for Biomedical Imaging, Department of Radiology, New York University School of Medicine, New York, NY 10016

Abstract. One overarching challenge of clinical magnetic resonance imaging (MRI) is to quantify tissue structure at the cellular scale of micrometers, based on an MRI acquisition with a millimeter resolution. Diffusion MRI (dMRI) provides the strongest sensitivity to the cellular structure. However, interpreting dMRI measurements has remained a highly ill-posed inverse problem. Here we propose a framework that resolves the above challenge for human white matter fibers, by unifying intra-voxel mesoscopic modeling with global fiber tractography. Our algorithm is based on a Simulated Annealing approach which simultaneously optimizes diffusion parameters and fiber locations. Each fiber carries its individual set of diffusion parameters which allows to link them by their structural relationships.

1 Introduction

Diffusion MRI (dMRI) has become an essential tool for noninvasive mapping of brain tissue [5]. A unique advantage of dMRI originates from the diffusion length, a typical displacement of water molecules, being of a few μm in a clinical scan, which is commensurate with cell dimensions. Hence, in addition to the millimeter-level anatomical MRI resolution, dMRI is sensitive to the tissue structure on the *mesoscopic scale* — an intermediate length scale between the molecular level where the NMR signal originates, and the macroscopic imaging voxel size. Mesoscopic brain tissue modeling attempts to quantify cellular-level tissue organization [2,1,4,8] in each voxel. However, determining the μm-scale parameters, such as axonal dimensions, water fraction and myelin thickness from a dMRI signal acquired at a 100–1000 times lower resolution, is a difficult problem riddled by unstable model fitting and sensitivity to noise. Fiber tractography, on the other hand, focuses on characterizing the structural connectome and inferring the interregional relationships of the human brain. It delineates white matter tracts based on the empirical anisotropy of the dMRI signal, and does not attempt to quantify the mesoscopic structure. Tractography algorithms come in many varieties, divided into deterministic streamline-based, probabilistic and global [7,9] approaches. Recently, the concept of tractometry [3] was developed, which projects mesostructural properties on top of already tracked fiber bundles.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 201–208, 2014.

This idea combines both fields above, however it is just a rather a retrospective 'combination' of outcomes of conceptually different approaches.

In contrast, this work tries to merge both fields into one framework such that both problems can benefit from each other. There are a few attempts that try similar approaches [6,10], but from an different perspective. The proposed algorithm performs a global tracking and local modeling simultanously. Instead of fitting the parameters independently in a voxel-by-voxel manner we treat the problem as one global optimization problem. On the one hand, we are able to link voxels by their structural relationships. On the other hand, the presence of fibers is purely driven by the fact that they correctly explain the observed signal and no further assumptions about topology are made. To make the optimization of this difficult, non-convex problem tractable we propose an efficient approximation of the data likelihood. A full brain reconstruction of the human brain takes about 12 hours on a standard Desktop PC.

2 The Fiber Model

The fiber model is built of small segments $X \in \mathcal{X}$. Each segment contributes to the predicted MR-signal $M(\mathcal{X})$ with a small signal contribution. Each segment carries its individual diffusion parameters that define this contribution. The segments can connect and polymerize to form long chains, called fibers. The set of edges connecting the segments is denoted by $\mathcal{E}$. The complete model $\mathcal{F} = (\mathcal{X}, \mathcal{E}, v)$ consists of the set of segments, their edges between them and the volume fractions v. Our mesoscopic model $M(\mathbf{r}, \mathbf{q})$ is composed of axially-symmetric Gaussian diffusion signals of the form

$$m_{\mathbf{n}}^{D_\parallel, D_\perp}(\mathbf{q}) = e^{-D_\parallel t(\mathbf{q}\cdot\mathbf{n})^2 - D_\perp t(|\mathbf{q}|^2 - (\mathbf{q}\cdot\mathbf{n})^2)}$$

from different white matter compartments (here t is a fixed diffusion time and the b-value [5] $b = |\mathbf{q}|^2 t$). The parameter $\mathbf{n}$ denotes the bundle direction, $D_\parallel$ and $D_\perp$ denote axial and radial diffusivities. The signal model is composed of the sum of two such tensor models, where for one of those the perpendicular diffusion is zero, and an additional constant reflecting non-diffusing water molecules. So, the signal from the ith segment is:

$$M_i(\mathbf{r}, \mathbf{q}) = \; v_r(\mathbf{r}) + m_{\mathbf{n}_i}^{D_\parallel^i, 0}(\mathbf{q}) v_a(\mathbf{r}) + m_{\mathbf{n}_i}^{D_\parallel^i, D_\perp^i}(\mathbf{q})(1 - v_a(\mathbf{r}) - v_r(\mathbf{r})). \qquad (1)$$

where v_r is the volume fraction of completely restricted water, v_a the intra-axonal fraction and $v_e = (1 - v_a - v_r)$ the extra-axonal fraction. While empirically this model captures the signal from a straight fiber bundle quite well, it currently has a limitation of setting the intra- and extra-axonal diffusivities to be the same. In general, these parameters seem not to be equal. The total expected signal is composed of a sum over all segments: $M(\mathbf{r}, \mathbf{q}) = \sum_{X_i \in \mathcal{X}} w_i \; I(\mathbf{r}, \mathbf{r}_i) M_i(\mathbf{q})$, where I is an indicator function giving contributions if $\mathbf{r}$ and $\mathbf{r}_i$ is in same voxel. Each segment carries 5 variables $X_i = (\mathbf{r}_i, \mathbf{n}_i, D_\parallel^i, D_\perp^i, w_i)$, the position, direction, the axial diffusivity along the fiber, the perpendicular diffusivity and its overall

weight. Note that the volume fractions $v_r(\mathbf{r})$, $v_a(\mathbf{r})$ and $v_e(\mathbf{r}) = 1 - v_a(\mathbf{r}) - v_r(\mathbf{r})$ are not properties of the segment but of the position. This is, on the one hand, conceptually quite natural and, on the other hand, a way to avoid the ambiguities of the model. To increase the number of segments (to get a higher number of fibers) the voxels may be divided into subvoxels which all share the same signal.

The cost functional, or energies as called in the following, consist of two parts: the data-likelihood and the prior that regularize the problem and control the connections between the segments. For optimization we use, like in [9], a simulated annealing approach. The idea is to simulate the Gibbs distribution $P(\mathcal{F}) = \frac{1}{Z}\exp(-(E_{\text{data}}(M(\mathcal{X}, v)) + E_{\text{prior}}(\mathcal{F}))/T)$ while lowering the temperature T. For lower temperature it gets more and more likely that we sample from minimum of the energy. The simulation principle is based on a Reversible Jump Monte Carlo Markov Chain (RJMCMC).

2.1 The Energy: Data Likelihood and Priors

The data term consists of a simple quadratic difference between signal and model, that is, we falsely assume a Gaussian data likelihood, which might cause a substantial bias on the parameters. However, we found in the numerical experiments that the Rician noise floor is mostly disrupting the v_r-fraction leaving the rest of the parameters nearly unbiased.

The priors control the number of segments, their connections, foster smoothness of the variables along fibers. Due to the freedom of the diffusion parameters we need a prior to prevent the fiber model to build unreasonable, non fiber like configurations, therefore we introduced an additional term E_{guide} similar to original data-likelihood, but each segment has a fixed diffusion model. We found that very sharp diffusion models, i.e. no extra-axonal compartment and high parallel diffusion, help to resolve sharp crossings. The second prior controls the number of particles and the third the number of connections. To each particle a cost is assigned, called chemical potential $E_{\text{chem}}(\mathcal{X}) = \mu|\mathcal{X}|$ where μ is strength of the prior, or equivalently the cost of one particle. The prior controlling the connection is similar to [9], but with one important extension. Each segment X has to two ports that can make connections with other segments. The location of the port is $\mathbf{r} \pm \ell\mathbf{n}$. If two segments are connected an additional potential is turned on which controls, the curvature and the similarity of the diffusion parameters. Let the segments X_1 and X_2 be connected, then we have the additional energy

$$U_{\text{con}}(X_1^{\alpha_1}, X_2^{\alpha_2}) = \lambda_d \sum_{P \in \{D_\parallel, D_\perp, v\}} (P^1 - P^2)^2 + U_{\text{bend}}(X_1^{\alpha_1}, X_2^{\alpha_2}),$$

where α_1, α_2 specify the ports. For a detailed description of the second term U_{bend} consult [9]. The first term gives an additional penalty on differences between the diffusion parameters, i.e. it drives the diffusion parameters to be similar along connected segments.

2.2 Approximation of Q-Space Correlations

The RJMCMC algorithm needs to compute energy differences like $E_{\text{data}}(M + M_{\text{mod}}) - E_{\text{data}}(M)$. The computation is dominated by correlations of the current model M with the newly added or modified segment M_{mod}, and the correlation of segment M_{mod} with the signal. The spatial part of these correlations is trivial, however the q-space part is quite costly as it involves the evaluation of the exponential model. To compute these correlations efficiently we found a power series approximations that can speed up the computation by an order of magnitude. The approximations are of type

$$\langle m_{\mathbf{n}}^{D_\parallel, D_\perp}, S\rangle_Q = \frac{1}{Q}\sum_{k=1}^{Q} m_{\mathbf{n}}^{D_\parallel, D_\perp}(\mathbf{q}_k)\; S(\mathbf{q}_k) \approx \sum_{l,m=1}^{M} \frac{b_{lm}(\mathbf{n})}{(\kappa + D_\parallel)^l (\kappa + D_\perp)^m} \quad (2)$$

where the $b_{lm}(\mathbf{n})$ can be found by a least squares minimization and the parameter κ is fixed and has to be found empirically to obtain good fits. The form is reminiscent of the Laplace transformation of exponential-type functions. For the two-shell scheme (a b=1000 and b=2000 shell) considered in the experiments we found $\kappa = 4$ to work well. We found values $M > 6$ do not improve fitting accuracy.

3 The Algorithm

As already stated the optimization of the proposed energies is accomplished by an RJMCMC-type algorithm together with a cooling process. The idea behind the RJMCMC-algorithm is to repeatedly make random distortions to the current state $\mathcal{F}$. The distortion, called $\mathcal{F}'$, usually depends on the previous state and follows some distribution $P_{\text{prop}}(\mathcal{F} \mapsto \mathcal{F}')$, which can be arbitrarily chosen by the algorithm designer. The only condition is that the reverse transition has to be possible, i.e. $P_{\text{prop}}(\mathcal{F}' \mapsto \mathcal{F}) > 0$. The algorithm needs usually a certain number of initial iterations such that the sequence of generated states follows the desired distribution and is in equilibrium. Once equilibrium is reached (which can be checked by statistics of the energy differences), the system is slowly cooled down. In the following we present the different proposals used in our implementation.

Segment Birth: A segment $X = (\mathbf{r}, \mathbf{n}, D_\parallel, D_\perp, w)$ is proposed by choosing all parameters uniformly. Then, the energy difference regarding the data-likelihood is computed according to $\Delta E_{\text{data}} = -2\langle M_X, S(\mathbf{r})\rangle_Q + 2\sum_k \langle M_X, M_k\rangle_Q + \langle M_X, M_X\rangle_Q$, where the sum over k ranges over all segments that lie within the voxel containing the new segment. For the efficient computation of such correlation the approximation (2) is used. The computation ΔE_{guide} is similar. Finally, the Gibbs ratio is $R = N_0 \exp(-(\Delta E_{\text{data}} + \Delta E_{\text{guide}})/T)/(N+1)$, where N is the number of segments currently present and N_0 the expected number of segments of the underlying Poisson process. **Segment death:** A segment X is randomly chosen. The energy differences that have to be computed are just the negated differences from the birth proposal. The Gibbs ratio

is $R = N \exp(-(\Delta E_{\text{data}} + \Delta E_{\text{guide}})/T)/(N_0)$. **Segment move** A segment X is randomly chosen. The position and orientation is distorted by normally distributed random numbers, $\mathbf{r}' := \mathbf{r} + \sigma_s \eta$ and $\mathbf{n}' := \mathbf{n} + \sigma_n \eta$. The Gibbs ratio is just $R = \exp(-(\Delta E_{\text{data}} + \Delta E_{\text{guide}} + \Delta E_{\text{con}})/T)$. **Change of segment's diffusion parameter** A segment X is randomly chosen. The current diffusion parameters are distorted by normally distributed random numbers, where the variance is proportional to the current temperature. The energy difference is computed in the same way like for the move proposal. **Change of volume fraction** A random voxel is chosen. Let us call E_M the data energy before the parameter change, then: $E_M = \sum_{k,j} \langle M_k, M_j \rangle_Q - 2 \sum_k \langle M_k, S(\mathbf{r}) \rangle_Q$, where the sum runs over all segments within the voxel. And correspondingly $E_{M'}$ after the change, then $\Delta E_{\text{data}} = E_{M'} - E_M$. In the same way like for the diffusion parameters the new volume fraction is proposed by distorting the old one by a normal distribution with a variance proportional to the current temperature. **Dis/Connecting segments** For the connection of segments follows the same principle as proposed in [9].

Parameters: The segment parameters are chosen similar to [9]. The length ℓ is chosen to be $2mm$ and the potential of connection is $L = 0.5$ (see [9] for notations). The chemical potential of a segment is chosen proportional to the number of measurement in q-space. We found $\mu = 0.005\ Q$ to be a good choice. That is, if a segment explains on average more than 0.005 of the variance of the signal, the segment is probably kept. For the strength of E_{guide} we found $\lambda_{\text{guide}} = 15T/T_{\text{start}}$ to work well. For the strength of the connection priors we found that values of $\lambda_c = 1$ and $\lambda_d = 1$ work already quite well. The temperature schedule starts at $T_{\text{start}} = 0.3$ and cools down to $T_{\text{end}} = 0.0025$, which corresponds to a SNR level of $1/\sqrt{T_{\text{end}}} = 20$.

4 Experiments

We consider a 2-shell scheme at b-values of 1000 and 2000 acquired with 60 directions per shell. The in vivo diffusion measurement was acquired on a Siemens 3T TIM Trio using an SE EPI sequence, with a TE of 107 ms. A healthy male volunteer (aged 36) was scanned at an isotropic resolution of 2.5mm. Additionally, a T_1 data set was acquired which was segmented into white matter (WM), gray matter (GM), and CSF using SPM. White matter was thresholded at a probability of 0.5 to determine the area of reconstruction.

First, to understand the importance of our approximation we did a brute force search on a synthetic data. By sweeping through the 3-parameter space of $D_{\parallel}, D_{\perp}$ and v_i we found that our approximation speeds up the likelihood computation by a factor of 20 compared to an ordinary implementation. To validate the accuracy of the approximation we simulated a simple crossing/bending configuration (see Figure 1) consisting of three bundles. The central crossing has a crossing angle of 50°. The phantom is simulated on $24 \times 24 \times 9$ grid with a isotropic voxel size of 2mm. Each of the three bundles has the same axonal volume fraction of 0.4, extra-axonal fraction of 0.6 and different diffusion parameters $(D_{\parallel}, D_{\perp})$. Bundle a) has $(1, 0.5)$, bundle b) $(1.5, 0.5)$ and bundle c) $(2, 1)$.

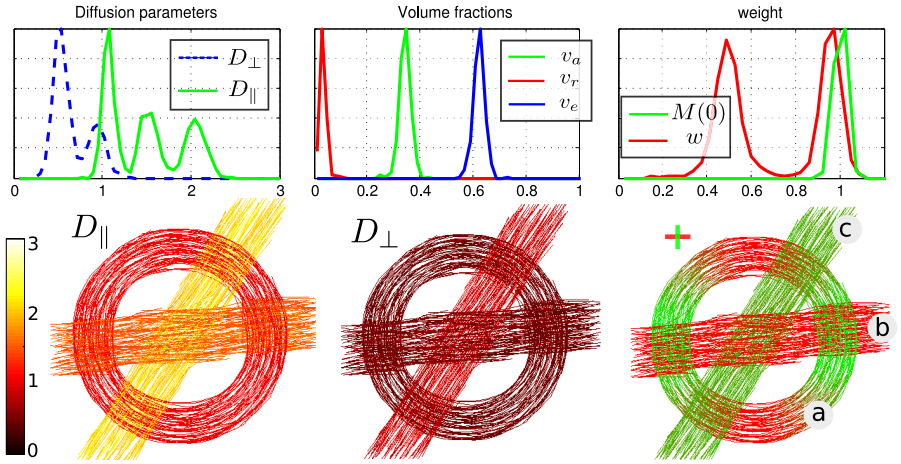


Fig. 1. Results for the phantom. Top: histograms of the diffusion parameters, volume fractions and signal weights/magnitudes. Bottom: Tracking result in direction coloring and coloring by their diffusion coefficients.

Rician noise was added with $\sigma = 0.05$ corresponding to a SNR of 20. Figure 1 shows histograms of our tracking results: fitted diffusion parameters, volume fractions, weight parameters and the tractogram. The reconstructed tracts are shown in three different coloring, one by directions, one by parallel diffusion $D_\parallel$ and one by perpendicular diffusion $D_\perp$. One can observe that all parameters are nearly unbiased. While the intra axonal volume fraction v_a shows a small underestimation, the v_r fractions and the diffusion coefficients show a small overestimation. For the in vivo dataset a voxel was subdivided into $3^3 = 27$ subvoxels to get a sufficient number of segments/fibers. For this setting the running time of the complete tracking procedure took about 10 hours on a Intel I7 (16GB) with four threads in parallel. The reconstruction contains 1.5 million particles forming about 50000 fibers longer than 10 segments.

In Figure 2 we show the results: Parametric brain maps of the diffusion parameters (Fig.2a), first and second order statistics (Fig.2b) of all parameters including tortuosity $t = D_\parallel/D_\perp$. The diagonal of the plot matrix shows ordinary histograms, the off-diagonal plots joint histograms of all parameter pairs. We also show histograms of the w parameter, the predicted signal at $b = 0$ and number of segments per voxels. Further, we selected several tracts (Fig. 2c) and d)) by two ROIs, namely, Cingulum (CG), Arcquate Fascicle (AF), Cortical Spinal Tract (CST), left Optic Radiation (OR), Fronto Occipital Fascicle (IFO) and callosal fibers to the precentral gyrus (CC). Finally, Figure 2e) shows fibers sliced coronally and colored by $D_\parallel$.

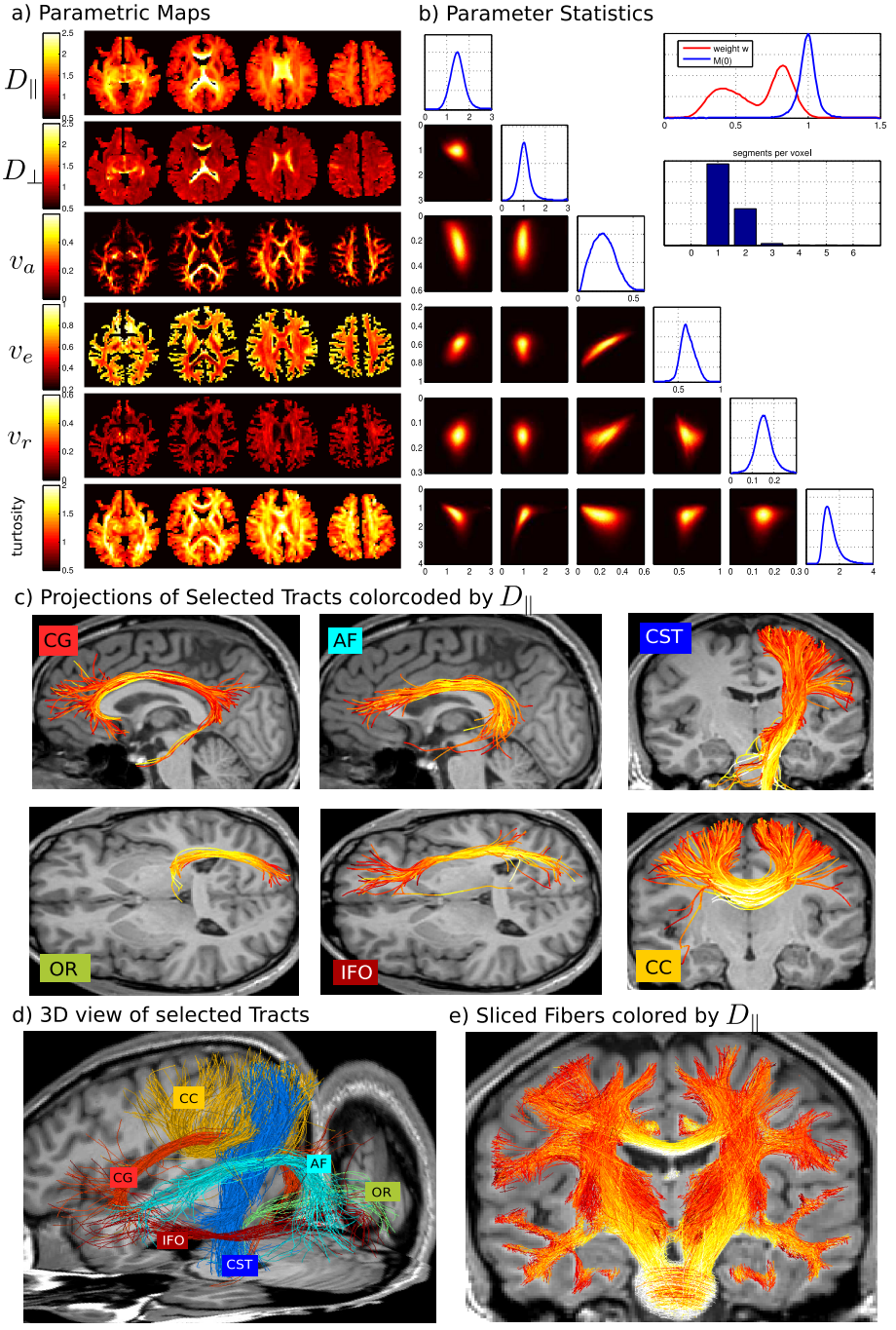


Fig. 2. In vivo-results for a 2-shell scheme (at b=1000,2000) at resolution of $2.5mm^3$

5 Discussion and Conclusion

We proposed a novel algorithm that unifies tractography and mesoscopic modeling to simultaneously reconstruct the human brain fiber bundle network and derives fiber specific diffusion parameters. The in vivo experiments show that the derived parameters go in-line with the current literature [4]. However, for the first time, we provide whole brain maps of the parameters including crossing regions. For single fiber voxel populations (like the Corpus Callosum) the putative axonal volume fraction (v_a) is in a range of about 40 to 50 percent, while $D_{\parallel} \approx 2$ and $D_{\perp} \approx 1$ which is similar to [4], where these parameters where derived via kurtosis imaging. The inferred parameters from multi fiber voxels differ, one can observe reduced v_a and $D_{\parallel}$ while an increase in $D_{\perp}$. The source of the restricted fraction v_r is not yet clear. There is definitely a certain amount caused by the Rician noise. The generated tract bundles show similar appearance like [9], but they additionally carry individual diffusion parameters.

Acknowledgement. The work of M. Reisert is supported by the Deutsche Forschungsgemeinschaft (DFG), grant RE 3286/2-1.

References

1. Alexander, D.C., Hubbard, P.L., Hall, M.G., Moore, E.A., Ptito, M., Parker, G.J.M., Dyrby, T.B.: Orientationally invariant indices of axon diameter and density from diffusion mri. NeuroImage 52, 1374–1389 (2010)
2. Assaf, Y., Blumenfeld-Katzir, T., Yovel, Y., Basser, P.J.: Axcaliber: a method for measuring axon diameter distribution from diffusion mri. Magn. Reson. Med. 59, 1347–1354 (2008)
3. Bells, S., Cercignani, M., Deoni, S., Assaf, Y., Pasternak, O., Evans, C.J., Leemans, A., Jones, D.K.: Tractometry–comprehensive multi-modal quantitative assessment of white matter along specific tracts. In: Proc. ISMRM, vol. 678 (2011)
4. Fieremans, E., Jensen, J.H., Helpern, J.A.: White matter characterization with diffusional kurtosis imaging. Neuroimage 58(1), 177–188 (2011)
5. Jones, D.K. (ed.): Diffusion MRI: Theory, Methods and Applications. Oxford University Press (2010)
6. Malcolm, J.G., Shenton, M.E., Rathi, Y.: Filtered multitensor tractography. IEEE Transactions on Medical Imaging 29(9), 1664–1675 (2010)
7. Mangin, J.: A framework based on spin glass models for the inference of anatomical connectivity from diffusion-weighted mr data - a technical review. NMR Biomed. 15(7-8), 481–492 (2002)
8. Panagiotaki, E., Schneider, T., Siow, B., Hall, M.G., Lythgoe, M.F., Alexander, D.C.: Compartment models of the diffusion MR signal in brain white matter: a taxonomy and comparison. Neuroimage 59(3), 2241–2254 (2012)
9. Reisert, M., Mader, I., Anastasopoulos, C., Weigel, M., Schnell, S., Kiselev, V.: Global fiber reconstruction becomes practical. Neuroimage 54(2), 955–962 (2011)
10. Sherbondy, A.J., Rowe, M.C., Alexander, D.C.: Microtrack: an algorithm for concurrent projectome and microstructure estimation. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part I. LNCS, vol. 6361, pp. 183–190. Springer, Heidelberg (2010)

Measurement Tensors in Diffusion MRI: Generalizing the Concept of Diffusion Encoding

Carl-Fredrik Westin[1,2], Filip Szczepankiewicz[3], Ofer Pasternak[1], Evren Özarslan[1], Daniel Topgaard[4], Hans Knutsson[2], and Markus Nilsson[3]

[1] Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[2] Department of Biomedical Engineering, Linköping University, Linköping, Sweden
[3] Lund University Bioimaging Center, Lund University, Lund, Sweden
[4] Center for Chemistry and Chemical Engineering, Lund University, Lund, Sweden

Abstract. In traditional diffusion MRI, short pulsed field gradients (PFG) are used for the diffusion encoding. The standard Stejskal-Tanner sequence uses one single pair of such gradients, known as single-PFG (sPFG). In this work we describe how trajectories in q-space can be used for diffusion encoding. We discuss how such encoding enables the extension of the well-known scalar b-value to a tensor-valued entity we call the diffusion measurement tensor. The new measurements contain information about higher order diffusion propagator covariances not present in sPFG. As an example analysis, we use this new information to estimate a Gaussian distribution over diffusion tensors in each voxel, described by its mean (a diffusion tensor) and its covariance (a 4th order tensor).

1 Introduction

In diffusion MRI (dMRI), each millimeter-size voxel of the image contains encoded information on the micrometer-scale translational displacements of the water [1]. The vast majority of applications today focus on the simplest form of the original MRI diffusion experiment, implemented by the Stejskal-Tanner pulse sequence [2]. This sequence is based on a pair of short pulsed diffusion encoding gradients, which we will refer to as the single pulsed field gradient (sPFG) experiment. sPFG typically is used in diffusion tensor imaging (DTI), enabling popular measures such as the mean diffusion (apparent diffusion coefficient, ADC) and diffusion anisotropy (Fractional Anisotropy, FA). Although current popular diffusion measures are very sensitive to changes in the cellular architecture, they are not very specific regarding the type of change.

We are at the cusp of a completely new generation of diffusion MRI technologies, such as oscillating gradients [3], double pulsed-field gradient (dPFG) sequences [4–6], and more general waveform sequences [7]. These methods are transforming what is possible to measure, and have the potential to vastly improve tissue characterization using diffusion MRI. Our work adds to this new generation of non-conventional pulse sequences. Our method can probe features of micron-scale transport processes (and thus microstructure) that are invisible with sPFG. Fig. 1 shows three example structures (voxel distributions) that

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 209–216, 2014.

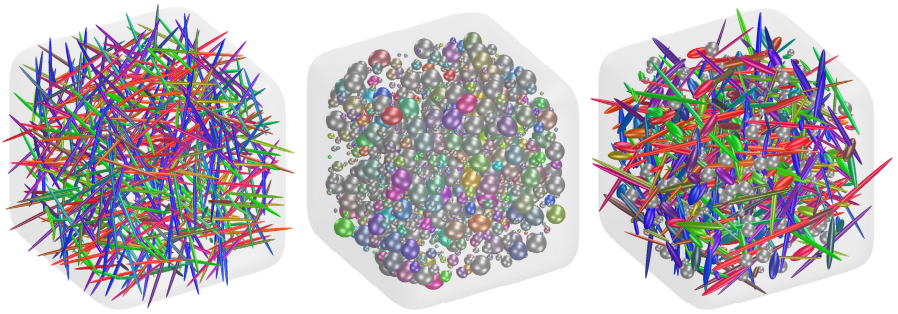


Fig. 1. Examples of globally isotropic distributions of structures within a voxel. These different structures are indistinguishable with traditional sPFG diffusion MRI.

would be indistinguishable using DTI. The aim of our work is the development of methods that can clearly distinguish these types of very different tissue architectures in clinical dMRI. In this paper, we present a new diffusion measurement framework and an example framework for analysis of the data we acquire. Together, these contributions enable us to quantify and distinguish distributions such as those in Fig. 1.

2 Theory

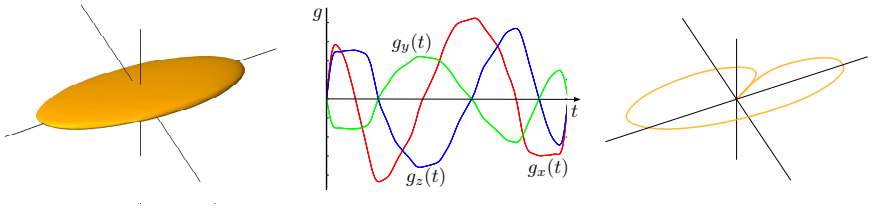
In conventional pulsed field gradient diffusion MRI, the diffusion encoding is achieved by applying a pair of short gradient pulses separated by a diffusion time. Such a measurement probes along a single axis in q-space. Here we will explore more general scenarios with time-varying gradients that probe trajectories in q-space. The geometry of the diffusion encoding can in the Gaussian approximation regime be described by a diffusion "measurement tensor," or "encoding tensor," which extends the traditional b-value to a tensor-valued entity. Here we define this measurement tensor by

$$\mathbf{B} = \int_0^\tau \mathbf{q}(t)\mathbf{q}^{\mathrm{T}}(t)\, dt\ , \quad \text{where} \quad \mathbf{q}(t) = \gamma \int_0^t \mathbf{g}(t')dt' \tag{1}$$

where $\mathbf{g}(t)$ is the time-dependent gradient, τ is the echo time, and γ is the gyromagnetic ratio. In this general case when the q-vector is built up by a time-dependent gradient to traverse an arbitrary path in q-space, the rank of the diffusion encoding tensor depends on the path, and is 1 in the case of sPFG, 2 for double-PFG, and 3 in the isotropic encoding case such as the triple-PFG [8] or q-MAS [9]. The conventional b-value is given by $b = \mathrm{Tr}(\mathbf{B})$, the trace of $\mathbf{B}$.

For example, a planar diffusion encoding tensor, i.e. an encoding that is rotationally symmetric in the plane (Fig. 2, left), can be achieved by a set of time

varying gradients (middle) that produce a planar q-space trajectory (right). Ideal planar encoding could be produced by a circular path in q-space. However, q-space encoding inevitably starts at the origin of q-space, so the path in Fig. 2 (right) is one way to obtain the planar encoding in practice. Constant angular b-value encoding can be ensured by varying the speed of the traversal in q-space, by using slower speed at low q-values, since the b-value is a function of both time and q-value. At a low q, a long diffusion time can build up the same encoding power (b-value), as a higher q-value with a shorter diffusion time.



a) measurement tensor b) x-,y-,z- gradients c) q-space trajectory

Fig. 2. An example of time varying gradients (a) that produce a q-space trajectory (b) and a planar measurement tensor in b-value encoding space (c)

To generate measurement tensors $\mathbf{B}$ with general shapes one can start with q-space trajectory $\mathbf{q}_0(t)$ that produces a diffusion measurement tensor $\mathbf{B}_0 = \int_0^\tau \mathbf{q}_0(t)\mathbf{q}_0(t)^{\mathrm{T}}\, dt$ and scale the trajectory with an affine transform $\mathbf{M}$ yielding the new curve $\mathbf{q}(t) = \mathbf{M}\mathbf{q}_0(t)$. This results in a new diffusion measurement tensor $\mathbf{B}$,

$$\mathbf{B} = \int_0^\tau \mathbf{M}\mathbf{q}_0(t)\ \left(\mathbf{M}\mathbf{q}_0(t)\right)^{\mathrm{T}}\ dt \tag{2}$$

$$= \mathbf{M}\ \left(\int_0^\tau \mathbf{q}_0(t)\mathbf{q}_0(t)^{\mathrm{T}}\ dt\right)\ \mathbf{M}^{\mathrm{T}} = \mathbf{M}\ \mathbf{B}_0\ \mathbf{M}^{\mathrm{T}} \tag{3}$$

The special case of transforming a normalized isotropic curve, $\mathbf{B_0} = \mathbf{I}$, produces the simple relation $\mathbf{B} = \mathbf{M}\mathbf{M}^T$ between the affine transform and the resulting measurement tensor.

We denote dMRI with encoding performed using arbitrary trajectories of $\mathbf{q}(t)$ as q-space trajectory imaging (QTI). The measurement tensors allowed by QTI enable the separation of orientation dispersion and underlying macroscopical dispersion [10–12]. Below we propose an example analysis of QTI data where we estimate a distribution over diffusion tensors at each voxel.

3 Methods

We implemented q-space trajectory imaging (QTI) on a clinical MRI scanner (Philips Achieva 3T). Imaging parameters were: TE = 160 ms, $\mathrm{Tr}(\mathbf{B}) = b = 0$,

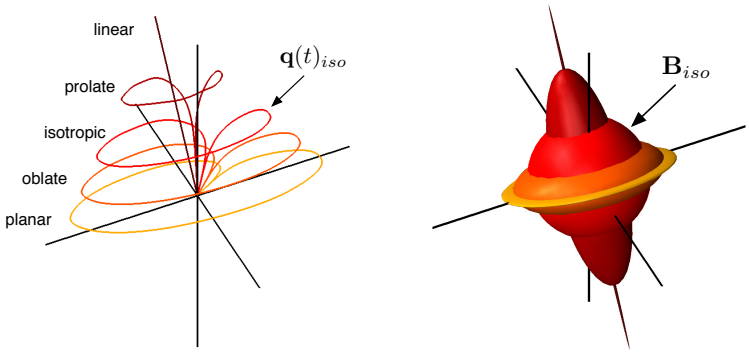


Fig. 3. Plot of five different q-space trajectories, with x-y-z axes (left). By varying the trajectory of q, diffusion encoding tensors of varying shapes can be produced. The color coding links the q-space trajectories (left) with the corresponding measurement tensors (right). The curve $\mathbf{q}(t)_{iso}$ produces a spherical b-value encoding.

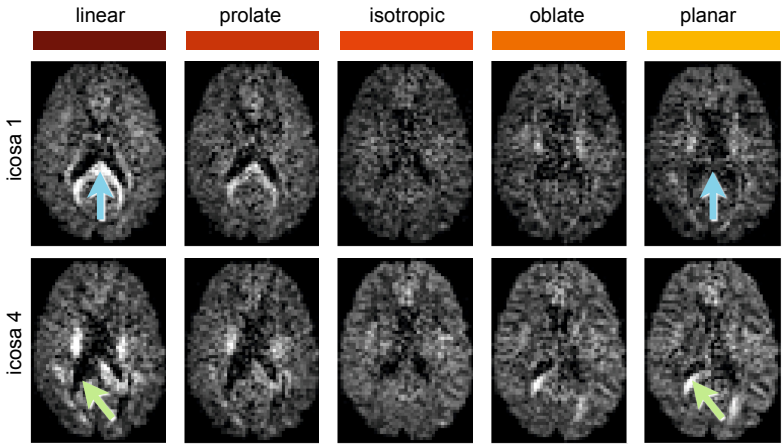


Fig. 4. MR signal from the five different types of trajectories in Fig 3, applied in two different directions in q-space. The five types of trajectories produce measurement **B**-tensors with (from left to right) linear, prolate, isotropic, oblate, and planar shapes. Note that the linear and the planar measurement are orthogonal/dual, and thus, where the linear measurement is bright the planar is dark; see blue and green arrows.

250, 500, 1000 and 2000 s/mm^2, voxel size = $3 \times 3 \times 3$ mm^3. The time varying gradients were designed to produce q-space trajectories generating linear, prolate, isotropic, oblate, and planar diffusion measurement tensors, which all were cylindrically symmetric, with the symmetry axis rotated into directions specified by the icosahedron, dodecahedron, and the truncated icosahedron. Despite the rather long TE due to our prototype implementation, the resulting diffusion encoded images were of a high image quality (Fig. 4).

3.1 Example Analysis: Estimating a Distribution over Diffusion Tensors

We propose an example analysis to demonstrate that we can measure additional microstructure information using QTI. Consider a system composed by a collection of environments, where in each environment the diffusion is Gaussian and described by the diffusion tensor $\mathbf{D}$ (as in Fig. 1). We propose to compactly model these microenvironments within a voxel with a Gaussian distribution over tensors. The tensor $\mathbf{D}$ is then a stochastic variable with expectation $\overline{\mathbf{D}} = \langle \mathbf{D} \rangle$. The covariance of $\mathbf{D}$ is given by a 4th-order tensor Σ of size $3 \times 3 \times 3 \times 3$ [13]. The description is simplified by using Voigt notation, which allows the diffusion tensor, which is normally expressed as a 3×3 matrix,

$$\mathbf{D} = \begin{pmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{pmatrix} \tag{4}$$

to be represented as a 1×6 vector

$$\mathbf{d} = \begin{pmatrix} D_{xx} & D_{yy} & D_{zz} & \sqrt{2}D_{xy} & \sqrt{2}D_{xz} & \sqrt{2}D_{yz} \end{pmatrix}^{\mathrm{T}} \tag{5}$$

allowing the fourth order 4th-order tensor Σ to be represented by a 6×6 variance-covariance matrix ($\mathbb{S}$), defined using the ordinary definition of the covariance matrix $\mathbb{S} = \langle \mathbf{d}\mathbf{d}^{\mathrm{T}} \rangle - \langle \mathbf{d} \rangle \langle \mathbf{d} \rangle^{\mathrm{T}}$, and in full given by

$$\mathbb{S} = \begin{pmatrix} \Sigma_{xxxx} & \Sigma_{xxyy} & \Sigma_{xxzz} & \sqrt{2}\Sigma_{xxxy} & \sqrt{2}\Sigma_{xxxz} & \sqrt{2}\Sigma_{xxyz} \\ \Sigma_{yyxx} & \Sigma_{yyyy} & \Sigma_{yyzz} & \sqrt{2}\Sigma_{yyxy} & \sqrt{2}\Sigma_{yyxz} & \sqrt{2}\Sigma_{yyyz} \\ \Sigma_{zzxx} & \Sigma_{zzyy} & \Sigma_{zzzz} & \sqrt{2}\Sigma_{zzxy} & \sqrt{2}\Sigma_{zzxz} & \sqrt{2}\Sigma_{zzyz} \\ \sqrt{2}\Sigma_{xyxx} & \sqrt{2}\Sigma_{xyyy} & \sqrt{2}\Sigma_{xyzz} & 2\Sigma_{xyxy} & 2\Sigma_{xyxz} & 2\Sigma_{xyyz} \\ \sqrt{2}\Sigma_{xzxx} & \sqrt{2}\Sigma_{xzyy} & \sqrt{2}\Sigma_{xzzz} & 2\Sigma_{xzxy} & 2\Sigma_{xzxz} & 2\Sigma_{xzyz} \\ \sqrt{2}\Sigma_{yzxx} & \sqrt{2}\Sigma_{yzyy} & \sqrt{2}\Sigma_{yzzz} & 2\Sigma_{yzxy} & 2\Sigma_{yzxz} & 2\Sigma_{yzyz} \end{pmatrix} \tag{6}$$

To estimate $\mathbb{S}$, consider the diffusion encoded MR-signal E from a system composed of multiple environments, each having Gaussian diffusion,

$$E(\mathbf{B}) = \left\langle \exp\left(- < \mathbf{B}, \mathbf{D} >\right) \right\rangle = \left\langle \exp\left(-\mathbf{b}^{\mathrm{T}}\mathbf{d}\right) \right\rangle \tag{7}$$

where $< \cdot \, , \, \cdot >$ is the inner product, which with Voigt notation is simplified to a vector inner product $< \mathbf{B}, \mathbf{D} >= \mathbf{b}^{\mathrm{T}}\mathbf{d}$ and $\left\langle \right\rangle$ represent integration over the distribution in the voxel. Expanding the logarithm of E around $\mathbf{B} = 0$ (derivation omitted), reveals a key relationship

$$\log E(\mathbf{b}) \approx -\mathbf{b}^{\mathrm{T}}\overline{\mathbf{d}} + \frac{1}{2}\mathbf{b}^{\mathrm{T}}\mathbb{S}\mathbf{b} \tag{8}$$

where $\overline{\mathbf{d}}$ is the mean value of $\mathbf{d}$. The equation is superficially similar to the model used in Diffusional Kurtosis Imaging (DKI), however, the fourth order kurtosis

tensor employed in sPFG DKI only has 15 unique elements in contrast to the 21 elements required to fully specify $\mathbb{S}$. QTI enables the probing of these 6 extra dimensions (21-15), not visible with sPFG. To estimate the covariance $\mathbb{S}$ (6x6 representation of the 4th-order tensor Σ) from a set of dMRI measurements, first note that

$$\mathbf{b}^{\mathrm{T}}\mathbb{S}\mathbf{b} =< \mathbf{b}\mathbf{b}^{\mathrm{T}}, \mathbb{S} >=< \mathbb{B}, \mathbb{S} >= \mathbb{b}^{\mathrm{T}}\mathbb{s} \tag{9}$$

where $\mathbb{b}$ and $\mathbb{s}$ are the Voigt representation of $\mathbb{B} = \mathbf{b}\mathbf{b}^{\mathrm{T}}$ and $\mathbb{S}$, as 21×1 vectors. Since Eq. 8 is a linear model, we may estimate $\overline{\mathbf{d}}$ and $\mathbb{s}$ using pseudoinversion to solve the following equation system

$$\begin{pmatrix} \log E_1 \\ \vdots \\ \log E_m \end{pmatrix} = \begin{pmatrix} 1 & -\mathbf{b}_1^{\mathrm{T}} & \frac{1}{2}\mathbb{b}_1^{\mathrm{T}} \\ \vdots & \vdots & \vdots \\ 1 & -\mathbf{b}_m^{\mathrm{T}} & \frac{1}{2}\mathbb{b}_m^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} E_0 & \overline{\mathbf{d}} & \mathbb{s} \end{pmatrix}^{\mathrm{T}} \tag{10}$$

In total, the model has 1+6+21 free parameters. $(E_0, \overline{\mathbf{d}}, \mathbb{s})$, The 21 parameters of the 4th-order tensor are difficult to interpret individually. The isotropic 4th-order tensor has, however, two components [14]

$$\mathbb{S}_{iso} = s_1\ \mathbb{E}_1 + s_2\ \mathbb{E}_2 \tag{11}$$

which in the field of mechanics are interpreted as bulk and shear modulus of the 4th-order stress tensor. The bases are given by

$$\mathbb{E}_1 = \frac{1}{3}\mathbf{e}\mathbf{e}^{\mathrm{T}} \quad \text{and} \quad \mathbb{E}_2 = \frac{3}{\sqrt{45}}(\mathbb{I} - \mathbb{E}_1) \tag{12}$$

where $\mathbb{I}$ is the 6×6 identity matrix. Note that $\mathbb{E}_1$ and $\mathbb{E}_2$ are orthogonal and normalized, i.e. $< \mathbb{E}_i, \mathbb{E}_j >= \delta_{ij}$. Expressed in full, these matrices assume simple structures according to

$$\mathbb{E}_1 = \frac{1}{3}\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \mathbb{E}_2 = \frac{1}{\sqrt{45}}\begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 \end{pmatrix} \tag{13}$$

Similarly to estimating the mean diffusivity MD by projecting the diffusion tensor on its isotropic basis element, $\mathbf{E} = \mathbf{I}/3$ (with $\mathbf{I}$ being the identity matrix), MD $=< \mathbf{D}, \mathbf{E} >$, we can project the estimated 4th-order covariance matrix onto its two isotropic basis elements $\mathbb{E}_1$ and $\mathbb{E}_2$ and obtain the parameters s_1 and s_2. These parameters can be interpreted as the bulk variation of diffusion tensors (i.e. variation in size) and the shear of them (i.e. variation between directions). Hence, s_2 contains information about microscopic anisotropy, and would give a high value for a system containing anisotropic microscopic compartments (Fig. 1, left), and a low value for isotropic compartments (Fig. 1, middle). On the other hand, s_1 reflects variation of mean diffusivities and would yield a low value if all microscopic compartments are similar in this respect (Fig. 1, left), but a high value if they are not (Fig. 1, middle).

s_1 (QTI) s_2 (QTI) FA (DTI)

Fig. 5. Bulk and shear modulus, s_1, s_2 estimated from fourth order QTI model, and FA from the DTI model

4 Results

Figure 5 shows the result of estimating the bulk and shear variation (s_1 and s_2) from QTI. The map of s_1 shows high values in regions where we expect both tissue and cerebrospinal fluid in the voxels, which leads to a high variability in mean diffusivities. By contrast, the map of s_2 is high and uniform in the white matter where variability in diffusivities is driven by the combination of high anisotropy and random orientations of the underlying microscopic environments. Since the analysis was performed on the isotropic components of the 4th order tensor, we know that all voxels have high orientation dispersion and thus s_1 reflects only underlying anisotropy. In contrast to FA from DTI, s_1 is high in regions of crossing fibers with a high orientation dispersion.

5 Discussion and Conclusions

QTI enables diffusion encoding with a general measurement tensor $\mathbf{B}$. Although the "b-matrix" concept is well established, and can be found in standard text books on diffusion NMR and MRI, the characterization of the b-matrix using double-PFG, and more general gradient wave form diffusion MRI is novel and different. In current literature, the concept of b-matrix normally refers to the standard rank-one measurement (in our terminology) with added imaging gradient and other correction terms. Extending the traditional rank-1 diffusion measurement, to rank-2 and full rank-3 measurements, allows for measuring information that was previously not attainable.

Our work shows that it is possible to perform diffusion encoding imaging of the human brain with arbitrary q-space trajectories while maintaining good SNR, and generalizes the concept of b-values enabling new types of measurements not available with sPFG.

Acknowledgments. The authors acknowledge the NIH grants R01MH074794, R01MH092862, P41-RR013218, P41EB015902, and the Swedish Research Council (VR) grants 2012-3682, 2011-5176, 2011-4334, and Swedish Foundation for Strategic Research grant AM13-0090.

References

1. Callaghan, P.T.: Translational dynamics and magnetic resonance: principles of pulsed gradient spin echo NMR. Oxford University Press (2011)
2. Stejskal, E.O., Tanner, J.E.: Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. J. Chem. Phys. 42(1), 288–292 (1965)
3. Does, M.D., Parsons, E.C., Gore, J.C.: Oscillating gradient measurements of water diffusion in normal and globally ischemic rat brain. MRM 49(2), 206–215 (2003)
4. Cory, D.G., Garriway, A.N., Miller, J.B.: Applications of spin transport as a probe of local geometry. Polym. Prepr. 31, 149–150 (1990)
5. Mitra, P.P.: Multiple wave-vector extensions of the NMR pulsed-field-gradient spin-echo diffusion measurement. Physical Review B 51(21), 15074 (1995)
6. Callaghan, P., Komlosh, M.: Locally anisotropic motion in a macroscopically isotropic system: displacement correlations measured using double pulsed gradient spin-echo NMR. Magnetic Resonance in Chemistry 40(13), S15–S19 (2002)
7. Drobnjak, I., Alexander, D.C.: Optimising time-varying gradient orientation for microstructure sensitivity in diffusion-weighted MR. JRM 212(2), 344–354 (2011)
8. Valette, J., Giraudeau, C., Marchadour, C., Djemai, B., Geffroy, F., Ghaly, M.A., Le Bihan, D., Hantraye, P., Lebon, V., Lethimonnier, F.: A new sequence for single-shot diffusion-weighted nmr spectroscopy by the trace of the diffusion tensor. MRM 68(6), 1705–1712 (2012)
9. Eriksson, S., Lasic, S., Topgaard, D.: Isotropic diffusion weighting in PGSE NMR by magic-angle spinning of the q-vector. JMR 226, 13–18 (2012)
10. Lasic, S., Szczepankiewicz, F., Eriksson, S., Nilsson, M., Topgaard, D.: Microanisotropy imaging: quantification of microscopic diffusion anisotropy and orientational order parameter by diffusion MRI with magic-angle spinning of the q-vector. Frontiers in Physics 2(11) (2014)
11. Jensen, J.H., Hui, E.S., Helpern, J.A.: Double-pulsed diffusional kurtosis imaging. NMR in Biomed. (2014)
12. Jespersen, S.S.N., Lundell, H., Snderby, C.K., Dyrby, T.B.: Orientationally invariant metrics of apparent compartment eccentricity from double pulsed field gradient diffusion experiments. NMR in Biomed. 26(12), 1647–1662 (2013)
13. Basser, P.J., Pajevic, S.: Spectral decomposition of a 4th-order covariance tensor: Applications to diffusion tensor MRI. Signal Processing 87(2), 220–236 (2007)
14. Moakher, M.: Fourth-order cartesian tensors: old and new facts, notions and applications. The Quart. Jour. of Mechanics and Applied Math. 61(2), 181–203 (2008)

From Expected Propagator Distribution to Optimal Q-space Sample Metric

Hans Knutsson[1] and Carl-Fredrik Westin[2]

[1] Linköping University, Sweden
[2] Harvard Medical School, USA

Abstract. We present a novel approach to determine a local q-space metric that is optimal from an information theoretic perspective with respect to the expected signal statistics. It should be noted that the approach does not attempt to optimize the quality of a pre-defined mathematical representation, the estimator. In contrast, our suggestion aims at obtaining the maximum amount of information without enforcing a particular feature representation.

Results for three significantly different average propagator distributions are presented. The results show that the optimal q-space metric has a strong dependence on the assumed distribution in the targeted tissue. In many practical cases educated guesses can be made regarding the average propagator distribution present. In such cases the presented analysis can produce a metric that is optimal with respect to this distribution. The metric will be different at different q-space locations and is defined by the amount of additional information that is obtained when adding a second sample at a given offset from a first sample. The intention is to use the obtained metric as a guide for the generation of specific efficient q-space sample distributions for the targeted tissue.

1 Introduction

The discussion concerning optimal q-space sampling strategies has been lively from the very start of diffusion imaging and is continuing to be a major topic of research [1] - [10]. Existing sampling schemes are based on experience combined with more or less *ad hoc* approaches of which many display interesting features. There is, however, no consensus regarding the choice of q-sample distribution in any given situation. Here we try to improve this situation by introducing a novel approach to determine a local q-space metric that is optimal from an information theoretic perspective with respect to the expected signal statistics.

The metric will be dependent on the q-space location an indicates the information gain, as a function of distance and direction, when adding a sample in a second q-space location. The obtained metric can then serve as a guide for the generation of specific q-space sample distributions e.q. sample distributions obtained in the manner described in [10]. It should be noted that the approach differs significantly from the classical estimation theory approach, e.g. one based on Cramer-Rao bounds [12]. The latter requires a pre defined mathematical

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 217–224, 2014.

representation, the estimator. Our suggestion aims at obtaining the maximum amount of information without enforcing a particular feature representation.

2 Theory

The mutual information (originally termed *rate of transmission*) between to signals relates directly to the entropies involved and can be estimated from the joint signal statistics [11]. Using a Gaussian signal+noise source model, which is a quite reasonable starting point in the present context, the estimate is directly related to the canonical correlation between two signals, a and b, and is given in bits by:

$$I_{ab} = -\frac{1}{2}\log_2\left(\frac{[\mathbf{C}_{aa}||\mathbf{C}_{bb}|}{|\mathbf{C}|}\right) \qquad \text{where} \qquad \mathbf{C} = \begin{pmatrix} \mathbf{C}_{aa} & \mathbf{C}_{ab} \\ \mathbf{C}_{ab} & \mathbf{C}_{bb} \end{pmatrix} \tag{1}$$

and $\mathbf{C}..$ are covariance matrices. For the one-dimensional case this reduces to: $I_{ab} = -\frac{1}{2}\log_2(1-\rho_{ab}^2)$, where ρ_{ab} is the correlation between the two variables. This expression can also be used to estimate the information from a single signal by measuring the correlation between the signal with and added noise realization and the same signal without noise. In order to obtain an estimate of a local information based q-space metric we can compute the information gain, I_Δ, from measuring in a second q-space location, q_b, given that we already have a measure at a first location, q_a. I_Δ is obtained as the information due to the second measurement alone minus the mutual information between the two measured signals, i.e. the information that is already present due to the first measurement:

$$I_\Delta \;=\; I_{b_0 b} - I_{ab} \;=\; \frac{1}{2}\log_2\left(\frac{1-\rho_{ab}^2}{1-\rho_{b_0 b}^2}\right) \tag{2}$$

In this expression b_0 is the true, noise free, source signal at the second location. It can be noted that equation (2) will give $I_\Delta = \frac{1}{2}\log_2\left(\frac{2\,\mathrm{SNR}+1}{\mathrm{SNR}+1}\right)$ if the second measurement is taken at the same location as the first. For reasonably high SNR (signal to noise ratio) this corresponds 0.5 bits or, equivalently, improving measurement SNR by $\sqrt{2}$ (3 dB).

It should also be noted that the Gaussian and additive assumptions are not crucial since mutual information between two variables is a monotonically increasing function of the correlation even in highly non-Gaussian and non-additive cases [13].

3 Method

To obtain the statistics of the q-space signals we generate a large number of q-space response examples. Using these examples correlation estimates between any two q-space locations, as well as correlations between different instances of

the same location, can be estimated. From these correlations the added information from measuring in a second q-space location, given a first measurement in any other location, can be found. The fact that each voxel in will contain a huge number of different propagators determining the q-space signals, and that a substantial intra voxel variation in propagator size and shape can be expected, makes it natural to use a Gaussian as a first approximation of the q-space response magnitude.

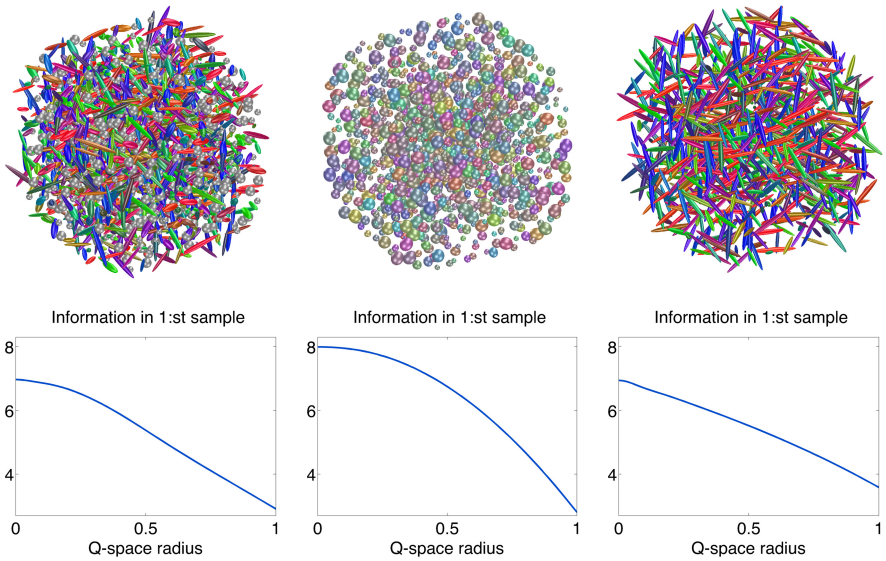


Fig. 1. The upper row illustrates three archetypal average propagator distributions. The plots show iso-surfaces of the Gaussian examples with centers distributed evenly in a volume of tissue. From left to right is shown: *Allsorts* - Varying in orientation, shape and size (Left). *Round* - Almost spherical propagators of varying in size (Center). *Stick* - Highly anisotropic only varying in orientation (Right). For each distribution the bottom row shows $I_{a_0 a}(\|\mathbf{q}\|)$, the amount of information that is given by the first sample at a given radius in q-space.

The example generator was set to produce 3D Gaussian q-space responses having one long axis and two equal short axes. All generated distributions had 300 different long axes orientations evenly distributed to cover all 3D orientations. The size of the average propagators was also varied. The total number of the propagator examples of a given 'tissue volume' was set to vary as the inverse of the volume, i.e. the total volume of the smaller propagators was equal to the total volume of the larger propagators. The average size of the propagators was set to vary logarithmically in the specified range. The ratio between the long and short axes was also set to vary logarithmically in the specified range while keeping the propagator volume constant. The intention is to study a few archetypal situations that can be easily understood, not to mimic real tissue. A

multitude of approaches for modelling biological tissue have been put forward and such models can readily be incorporated in the present framework. Doing so is, however, beyond the scope of this paper.

Listed below and shown in figure 1 are the three different distributions generated to study the effects different average diffusion propagator distributions will have on the q-space sampling metric. All three generated distributions are rotation invariant. Also note that the individual spatial propagator positions are not important for our analysis.

Allsorts - Long/short axis ratios from 1 to 10 and volumes from 0.5 to 2.
Round - A long/short axis ratio of 1.15 and volumes from 0.22 to 4.5.
Stick - A long/short axis ratio of 10 and volumes from 0.5 to 2.

The lower part of figure 1 shows the information given by the first sample as a function of q-space radius. The q-space radius refered to in the present work is a relative entity, the relation to physical q-space depends on scanner setup and actual noise levels. That the information decreases with q-space radius is a consequence of the diminishing average signal energy present for all three distributions (and for all reasonable other distributions).

4 Results

The *Allsorts* distribution: Figure 2 shows the result of the estimated q-space metric for the *Allsorts* distribution. The lilac colored iso-surfaces show the 3D q-space locations where the information gain from a second sample, given a first sample in the center (yellow), reaches $\Delta I = 2$ bits. Results for five different radii (0, 0.25, 0.5, 0.75 and 1.0) of the first sample are shown. The radii were chosen to highlight the typical information gain behaviors that will be present in different parts of q-space. Since the setup is rotationally invariant the results will be the same along any axis through the origin. The upper plot shows iso-surfaces along four different directions in one octant of q-space and is intended to demonstrate that the metric is rotation invariant. The lower part of figure 2 shows the five results obtained for the initial point located at different positions on the x-axis. A short summary of the situation at different q-space radii, r, is given below:

- **At r=0:** The second sample must be moved quite far from the first to gain more information, i.e. very sparse sampling is needed, a sample at q=0 picks up most of the information available at the center.
- **At r=0.25:** The information gain now quickly increases in an approximately isotropic fashion when the second location is moved away from the first. This indicates that a relatively dense sampling is preferable here.
- **At r=0.5:** The situation resembles the previous one but a slight anisotropy of the 2-bit iso-surface can be noted. The second sample has to be moved further in the radial direction than in an angular direction to give the same information gain.
- **At r=0.75:** The anisotropy is becoming more pronounced indicating that moving the second location in the angular direction is clearly preferable to

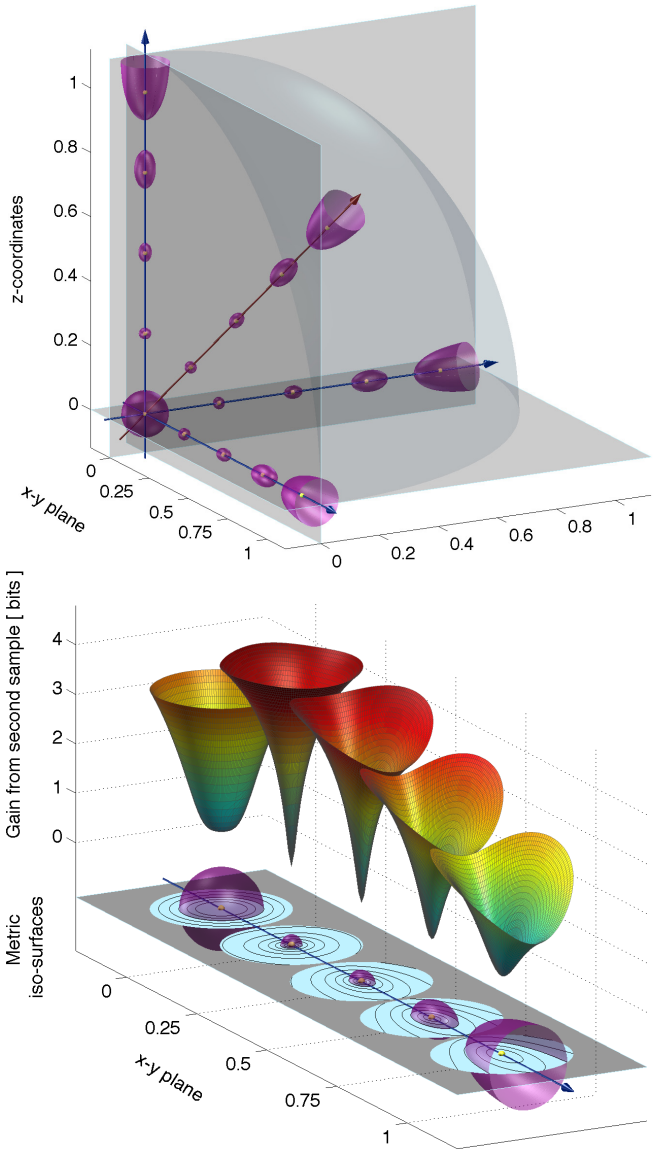


Fig. 2. Figure showing the result of the estimated q-space metric for the *Allsorts* distribution. The lilac colored iso-surfaces show where the information gain from a second sample, given a first sample in the center (yellow), reaches $\Delta I = 2$ bits. The figure shows the results for five different radii of the first sample. The upper plot shows one octant of q-space and is intended to demonstrate that the metric is rotation invariant. The lower plot shows only the cases where the first sample is on the x-axes. The multi colored surfaces show the information gain when moving the second sample away from the first, the gain in bits is indicated by the numbers 1-4 on the upper part of the y-axis. The iso-contours on the x-y plane below are drawn for every 0.5 bits.

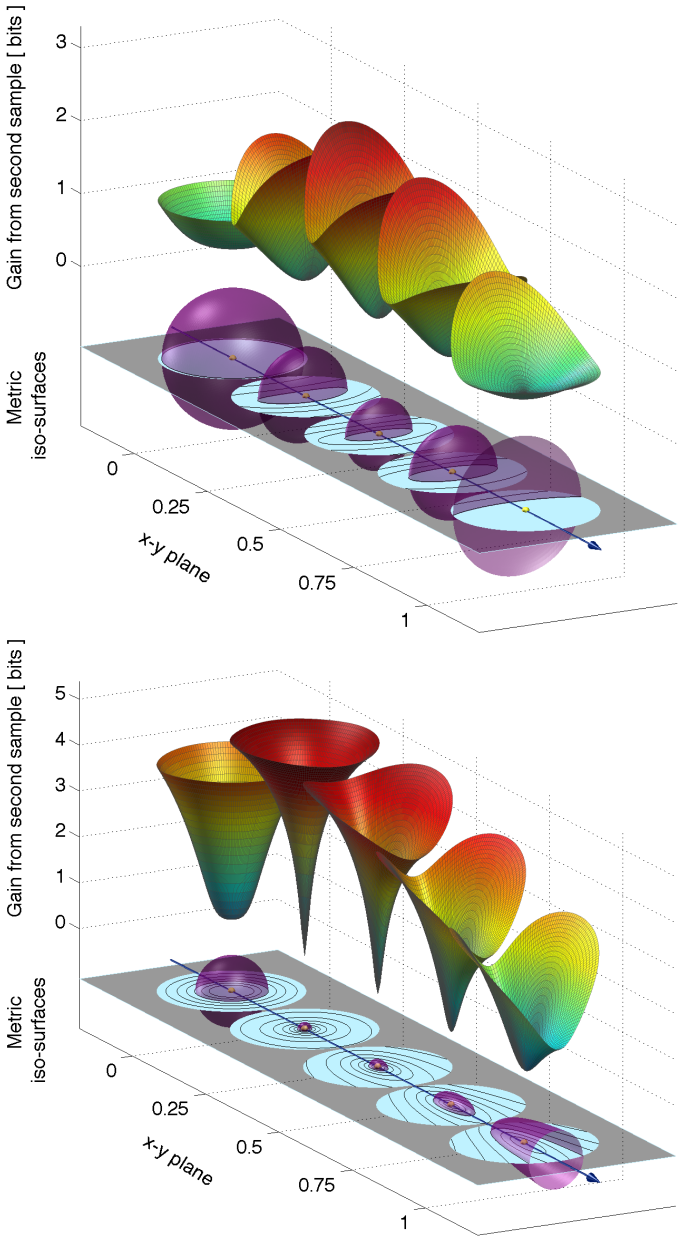


Fig. 3. Figure demonstrating the estimated q-space metric dependence on the propagator distribution. The upper plot shows the result for the *Round* distribution. The lilac colored iso-surfaces show where the information gain from a second sample, given a first sample in the center (yellow), reaches $\Delta I = 1$ bit. The lower plot shows the result for the *Stick* distribution. Here the iso-surfaces are drawn at $\Delta I = 2$ bits. For both plots the iso-contour line are drawn 0.5 bits apart.

a change in radius. The volume enclosed by the iso-surface is also larger indicating that a less dense sampling is needed. The sampling distance in the radial direction should be lower than that in the angular direction.
- **At r=1.0:** The iso-surface now takes the shape of a cone implying that the information gain is largest when moving simultaneously inwards and angularly. This is due to that the SNR at this radius quickly decreases with increasing radius due to loss of signal strength.

The *Round* and *Stick* Distributions: By studying distribution that are individually more uniform and 'removed' from the *Allsorts* distribution in two different ways further insights can be gained. The plot at the top of figure 3 shows the results when using the *Round* distribution and below the results from using the *Stick* distribution is plotted. Comparing the two results the following can be noted:

- **At r=0:** In both cases the second sample still must be moved relatively far from the first to gain more information. This effect is very pronounced for the *Round* distribution and the plots indicate that the second sample most be moved roughly four times as far as for the *Stick* distribution to gain the same amount of information. (Note that the iso-surfaces are drawn at different levels, $\Delta I = 1$ bit and $\Delta I = 2$ bits.)
- **At r=0.25:** Here the difference between the two distribution is even bigger. For the *Round* distribution very little is gained by by displacing the second sample angularly, the preferred displacement direction is clearly radial. In contrast the information gain increases very quickly in an isotropic fashion for the *Stick* distribution.
- **At r=0.5 and r=0.75:** The situation resembles the previous one but an increasing anisotropy of the 2-bit iso-surface can be noted for the *Stick* distribution. The second sample has to be moved further in the radial direction than in an angular direction to give the same information gain. A common feature is that the information gain vs displacement distance is decreasing for both distributions with increasing q-space radius.
- **At r=1.0:** The *Stick* distribution iso-surface now takes the shape of an open cone implying that the information gain is largest when moving simultaneously inwards and angularly. For the *Round* distribution the iso-surface almost becomes a plane. This indicates that we are close to a radius were higher information gain can only be obtained by moving towards the center, we are approaching the information 'edge' of q-space.

A general difference that is globally present is that it that the average information gain from a second sample is much lower for the *Round* distribution than for the *Stick* distribution. This is, most likely, a consequence of that the former distribution has a lower overall variability, i.e. lower entropy.

5 Discussion and Conclusion

Although the interpretation of the results may be accordance with the gut feeling of some experienced researchers in the field we believe that our analysis provides

a novel view allowing a quantification of said feeling. Our analysis also stresses the fact that the actual average propagator distribution in the targeted tissue has a major effect on what is the optimal q-space sampling strategy. In many cases tissue models can be employed and educated guesses can be made regarding the average propagator distribution present. In such cases the presented analysis can, for example, be used to find parameters for the 3D q-space sample distribution scheme described in [10]. In this way full q-space sampling, optimal with respect to a given expected distribution of diffusion propagators can be produced. This will also allow tuning of q-space distributions to maximize resolution for targeted tissue features.

Acknowledgement. The authors acknowledge the Swedish Research Council grants 2011-5176, 2012-3682, 2007-8632 (CADICS Linneaus research environment) and NIH grants R01MH074794, P41RR013218, and P41EB015902.

References

1. Jones, D.K., Simmons, A., Williams, S.C.R., Horsfield, M.A.: Non-invasive assessment of axonal fiber connectivity in the human brain via diffusion tensor MRI. Magn. Reson. Med. 42, 37–41 (1999)
2. Assaf, Y., Freidlin, R.Z., Rohde, G.K., Basser, P.J.: New modeling and experimental framework to characterize hindered and restricted water diffusion in brain white matter. Magn. Reson. Med. 52(5), 965–978 (2004)
3. Wu, Y.C., Alexander, A.: Hybrid diffusion imaging. Neuroimage 36(3), 617–629 (2007)
4. Alexander, D.C.: A general framework for experiment design in diffusion MRI and its application in measuring direct tissue-microstructure features. Magn. Reson. Med. 60(2), 439–448 (2008)
5. Merlet, S., Caruyer, E., Deriche, R.: Impact of radial and angular sampling on multiple shells acquisition in diffusion MRI. Med. Image Comput. Comput. Assist. Interv. 14(Pt. 2), 116–123 (2011); Eds: T. Peters, G. Fichtinger, A. Martel
6. Caruyer, E., Cheng, J., Lenglet, C., Sapiro, G., Jiang, T., Deriche, R.: Optimal Design of Multiple Q-shells experiments for Diffusion MRI. In: MICCAI Workshop CDMRI 2011 (2011)
7. Ye, W., Portnoy, S., Entezari, A., Blackband, S.J., Vemuri, B.C.: An Efficient Interlaced Multi-shell Sampling Scheme for Reconstruction of Diffusion Propagators. IEEE Trans. Med. Imaging 31(5), 1043–1050 (2012)
8. Westin, C.F., Pasternak, O., Knutsson, H.: Rotationally invariant gradient schemes for diffusion MRI. In: Proc. of the ISMRM Annual Meeting (ISMRM 2012), p. 3537 (2012)
9. Scherrer, B., Warfield, S.K.: Parametric Representation of Multiple White Matter Fascicles from Cube and Sphere Diffusion MRI. PLoS One 7(11), 1–20 (2012)
10. Knutsson, H., Westin, C.-F.: Charged Containers for Optimal 3D Q-space Sampling. In: ISMRM 2013 (2013)
11. Shannon, C.: Communication in the Presence of Noise. Proceedings of the IRE 37(1), 10–21 (1949)
12. Rao, C.R.: Bull. Calcutta Math. Soc. 37 (1945)
13. Pires, C.A.L., Perdigo, R.A.P.: Minimum Mutual Information and Non-Gaussianity Through the Maximum Entropy Method: Theory and Properties. Entropy 14, 1103–1126 (2012); Ed: K. H. Knuth

Image Quality Transfer via Random Forest Regression: Applications in Diffusion MRI

Daniel C. Alexander[1], Darko Zikic[2], Jiaying Zhang[1], Hui Zhang[1], and Antonio Criminisi[2]

[1] Centre for Medical Image Computing and Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, UK
[2] Microsoft Research Cambridge, Cambridge, UK

Abstract. This paper introduces image quality transfer. The aim is to learn the fine structural detail of medical images from high quality data sets acquired with long acquisition times or from bespoke devices and transfer that information to enhance lower quality data sets from standard acquisitions. We propose a framework for solving this problem using random forest regression to relate patches in the low-quality data set to voxel values in the high quality data set. Two examples in diffusion MRI demonstrate the idea. In both cases, we learn from the Human Connectome Project (HCP) data set, which uses an hour of acquisition time per subject, just for diffusion imaging, using custom built scanner hardware and rapid imaging techniques. The first example, super-resolution of diffusion tensor images (DTIs), enhances spatial resolution of standard data sets with information from the high-resolution HCP data. The second, parameter mapping, constructs neurite orientation density and dispersion imaging (NODDI) parameter maps, which usually require specialist data sets with two b-values, from standard single-shell high angular resolution diffusion imaging (HARDI) data sets with $b = 1000\,\mathrm{s\,mm^{-2}}$. Experiments quantify the improvement against alternative image reconstructions in comparison to ground truth from the HCP data set in both examples and demonstrate efficacy on a standard data set.

1 Introduction

Bespoke MRI scanners and imaging protocols can produce very high quality data uniquely informative about anatomy and function. However, the imaging techniques that underpin such data sets are often impossible or impractical on standard devices or in clinical imaging scenarios. For example, small-animal scanners often have higher field and gradient strength and smaller bore than human scanners, enhancing signal to noise, image resolution and, in diffusion MRI, sensitivity to small structures. Such platforms can provide exquisitely high resolution images revealing fine structural detail and providing strong sensitivity to anatomical features or pathology. Although such measurements highlight the potential of future human imaging devices, they provide little direct benefit to current clinic practice. Similarly, the HCP designed bespoke MRI scanners equipped

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 225–232, 2014.

with $100\,\mathrm{mT\,m^{-1}}$ and $300\,\mathrm{mT\,m^{-1}}$ gradient coils (several times more powerful than standard clinical scanners) and exploit several imaging and image reconstruction innovations to speed up acquisition and improve data quality [1]. The bespoke imaging system combined with a lengthy acquisition protocol leads to unique data sets with unprecedented image resolution and noise levels. However, the techniques extend only partially to clinical imaging with modest hardware and much more limited imaging times.

In this paper, we propose to exploit the information in expensive high quality data sets to improve images reconstructed from more modest data acquisitions. We call this process *image quality transfer*. We learn fine image structure from high quality data sets and use it to enhance lower quality data. We present a framework for solving this general problem using a patch-based image representation and random forest regression.

Two distinct applications demonstrate the framework by exploiting the HCP in-vivo human diffusion MRI data, which has uniquely high quality. The first application is super-resolution of DTIs. The HCP diffusion data have voxel sizes of $1.25^3\,\mathrm{mm^3}$ rather than typical sizes around $2^3\,\mathrm{mm^3}$ in standard data sets. Image quality transfer provides a mechanism to reconstruct high resolution DTIs from low-resolution acquisitions. The second application is quantitative parameter mapping. The HCP provides three HARDI shells of data with diffusion weighting factor $b \approx 1000$, 2000, and $3000\,\mathrm{s\,mm^{-2}}$, which supports estimation of more informative parameters than standard data sets with a single HARDI shell at $b = 1000\,\mathrm{s\,mm^{-2}}$. For example, NODDI [2] provides more specific information than DTI, such as maps of the density and dispersion of neurites (axons and dendrites), by fitting a more informative model in each voxel. It has become popular for clinical studies, because it requires as little as 15 minutes acquisition time. However, NODDI requires at least two HARDI shells with distinct b and fitting the NODDI model fails with only a single HARDI shell [2]. This prevents its use on the large variety of historical standard data sets. Image quality transfer provides a mechanism to recover NODDI parameter maps from single b-value data sets, which potentially enables NODDI analysis of historical data.

Prior literature on super-resolution is extensive. In medical imaging, [3] uses example patches from high resolution images to super-resolve scalar MR images and [4] use dictionaries from a database of similar images. Several authors propose super-resolution techniques specifically for diffusion images. The closest to our work is [5], which enhances the resolution of each diffusion weighted image (DWI) through patch examples before fitting the DT or other models; the discussion compares this approach to ours in more detail. Image quality transfer for parameter mapping from rarefied data sets holds greater novelty, although similar in spirit to modality transfer [6], which predicts T2 and FA images from T1 scans via patch-based label propagation. Our framework solves both problems.

2 Methods

Our implementation of image quality transfer learns a mapping from each neighbourhood of N_1 voxels in the low-quality data set to a corresponding

neighbourhood of N_2 voxels in the high-quality data set. Input and output voxels are vector-valued containing p_1 and p_2 values, respectively. Construction of the mapping is thus a regression problem. It requires a training set of patch pairs $T = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{|T|}$, where each $\mathbf{x}_i$ has dimension $p_1 N_1$ and $\mathbf{y}_i$ dimension $p_2 N_2$.

2.1 Regression Models

We consider a hierarchy of three types of mapping, where each generalises the previous: global linear; regression trees; and regression forests.

For global linear regression, we compute the linear transformation matrix $G = YX^\dagger$, where Y has columns $\mathbf{y}_i$, X has columns $\mathbf{x}_i$, and $X^\dagger$ is an appropriate pseudo inverse of X so that, in matlab, `G = X\Y`. For input patch $\mathbf{x}$, the estimate of the corresponding output patch is $G\mathbf{x}$.

The regression tree implements a piecewise linear regression over the space of input data points [7,8]. Each internal node in the tree sends data points into left or right subtrees by thresholding one of J scalar functions of $\mathbf{x}$, or *features*, $F_1, \cdots, F_J$. The choice of features is application dependent and we define ours later. Each leaf node contains a linear transformation with the same structure as the global linear transformation G defined in the previous section. Thus, for input data point $\mathbf{x}$, the output estimate is $G_t\mathbf{x}$ where G_t is the linear transformation of the leaf node at which the data point arrives after traversing the tree. Training uses a standard greedy search strategy similar to [8]. To control for overfitting, we use a validation set V with similar size to T and accept only splits that reduce the residual error of V.

Regression forests use multiple regression trees constructed from different training sets. Outputs are element by element averages of the prediction from each tree weighted by the error covariance of the linear transformation G_t, estimated during training.

2.2 Application 1: Super-Resolution

For DTI super-resolution, the mapping takes a $(2n+1) \times (2n+1) \times (2n+1)$ cubic patch of DTs, so that $N_1 = (2n+1)^3$ and $p_1 = 6$, as input, and outputs an $m \times m \times m$ cubic patch of voxels, each also containing a DT, so that $N_2 = m^3$ and $p_2 = 6$. The output patch is a cubic array of subvoxels that super-resolves the central voxel of the input patch. For example, if $n = 2$ and $m = 2$, each input $\mathbf{x}$ contains the 6 independent elements of the DT in each voxel of a $5 \times 5 \times 5$ low-resolution patch and each output $\mathbf{y}$ the elements of each DT in the $2 \times 2 \times 2$ high resolution patch (see figure 1); the mapping is thus from $\mathcal{R}^{750}$ to $\mathcal{R}^{48}$.

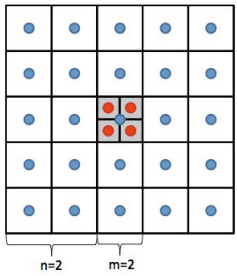


Fig. 1. 2D illustration of the input (blue) and output (red) patch structure for $n = 2$ and $m = 2$

For $F_1, \cdots, F_J$, we use the following features of $\mathbf{x}$:

– The three eigenvalues of the DT in the central voxel.

- The orientation of the principal eigenvector of the DT in the central voxel.
- The means of each eigenvalue over the central $3 \times 3 \times 3$ cube and those over the whole $(2n+1)^3$ cube.
- The mean principal orientation over the central 3^3 and the whole cubes.
- The orientational variance over the central 3^3 and the whole cubes.

Unless otherwise stated, training data comes from 8 randomly selected HCP data sets `www.humanconnectome.org`. A separate test set contains a different 8 HCP data sets. Each data set contains 288 DWIs including 18 with $b \approx 0$ and three HARDI shells of 90 directions with $b \approx 1000$, 2000, and $3000\,\mathrm{s\,mm^{-2}}$; the precise values vary spatially, as described in [1]. This application uses only the $b \approx 0$ and $b \approx 1000\,\mathrm{s\,mm^{-2}}$ measurements to reflect standard data sets. Training pairs come from downsampling each DWI by a factor of m in each dimension, fitting the DT in each voxel of both the downsampled and full resolution image using weighted linear least squares accounting for the spatially varying b and gradient directions, and associating $(2n+1)^3$ patches in the downsampled image with the m^3 patch in the full resolution image corresponding to the central voxel of the low-resolution patch.

Each data set contains around 7.5×10^5 brain voxels. We randomly subsample the set of patches from the full training set to meet memory limitations.

2.3 Application 2: Parameter Mapping

This application aims to estimate maps of NODDI parameters from standard data sets including only $b \approx 0$ and $b \approx 1000\,\mathrm{s\,mm^{-2}}$ measurements. Thus the mapping takes as input a $(2n+1)^3$ cubic patch of DTs fitted to a $b \approx 1000\,\mathrm{s\,mm^{-2}}$ HARDI shell, so that $N_1 = (2n+1)^3$ and $p_1 = 6$. The mapping outputs the NODDI parameters, intra-cellular volume fraction f_{ICVF}, free-water volume fraction f_{ISO}, orientation dispersion index (ODI), and the mean fibre orientation (θ, ϕ), at the central voxel of the input patch; $N_2 = 1$ and $p_2 = 5$.

The features, $F_1, \cdots, F_J$, training, and test data sets are as in DTI super-resolution. The ground truth NODDI output for the training set comes from fitting the NODDI model to all three HARDI shells in each image voxel.

3 Results

Figure 2 shows qualitative results from both applications on one of the test data sets (not used in training). The left panel compares various super-resolution reconstructions, obtained with $m = 4$, after downsampling by a factor of 4 in each dimension. The right panel of figure 2 compares various reconstructed NODDI maps with ground truth. In both examples, $n = 2$, the training set T contains about 1.5M data points, and the forest uses 8 trees.

For super-resolution, clarity in both the global linear and forest reconstructions compare favourably to standard interpolation techniques. The global linear and forest reconstructions appear quite similar, although the latter avoids some

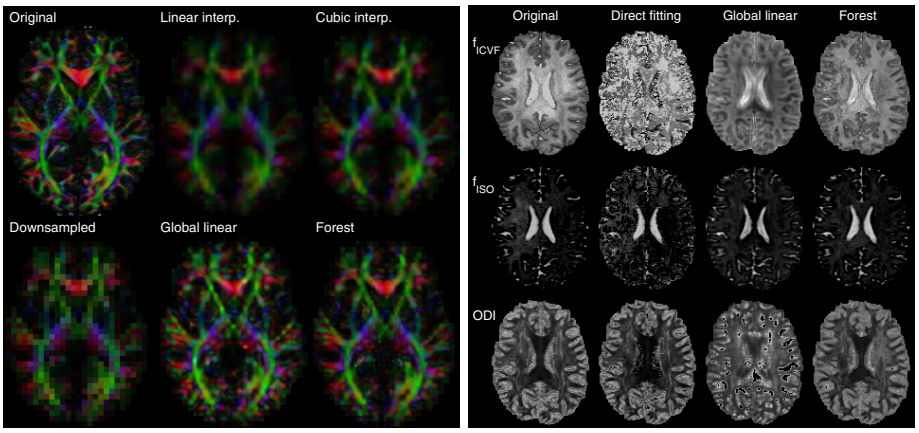


Fig. 2. Left: Direction-encoded colour FA maps for various reconstructed DTIs from a downsampled image (bottom left) compared to ground truth (top left) from the original full resolution data set. Right: Comparison of ground truth NODDI parameter maps (left) with standard fitting to just the $b \approx 1000\,\mathrm{s\,mm^{-2}}$ shell (left middle), global linear (right middle) and random forest (right) regression.

glitches visible in the former. As [2] predicts, the standard voxel-by-voxel NODDI parameter estimation fails, strongly disrupting the neurite density parameter, f_{ICVF}, in particular. The global linear transformation also performs poorly and fails to recover the structure of the f_{ICVF} and ODI maps, although it does produce a reasonable reconstruction of f_{ISO}. The improvement of the forest over the global linear transformation is striking and the output is visually much closer to the ground truth, although some differences are still clearly discernible.

Figure 3 quantifies the comparison of high resolution (left) and NODDI parameter (right) reconstructions as a function of training set size. The metric of reconstruction error is the mean (over the 8 test subjects) median (over brain voxels) root-mean-squared parameter error (i.e. of the six independent DT elements or 3 scalar NODDI parameters). All regression techniques improve on standard interpolation methods for super-resolution. Trees and forests improve on global linear transformations increasingly as the training set size increases, because they exploit additional training data by increasing the complexity of the mapping model. Averaging over the output of multiple trees (the forest) shows benefit over single trees. The advantage comes in part because the forests see more training data than single trees, as each component tree uses a different random training set, although they also mitigate the greedy search that trains individual trees. Available computer memory limits the amount of training data a single tree can use, so the forests offer genuine advantages by enabling exploitation of more training data.

Between subject standard deviation is consistent among the different algorithms suggesting that the error score depends on individual anatomical features,

such as the size of the ventricles where the error scores tend to be largest for all algorithms. Other DT metrics, such as orientational difference, show similar trends. The standard deviation over training set randomisation is more meaningful than over subjects for evaluating the significance of differences in error score. That standard deviation is 3 or 4 orders of magnitude smaller than the error itself in most cases, which suggests that (a) the results are highly reproducible over different training sets; and (b) the differences in error scores between, e.g. trees and forests, are highly significant (around 60 standard deviations).

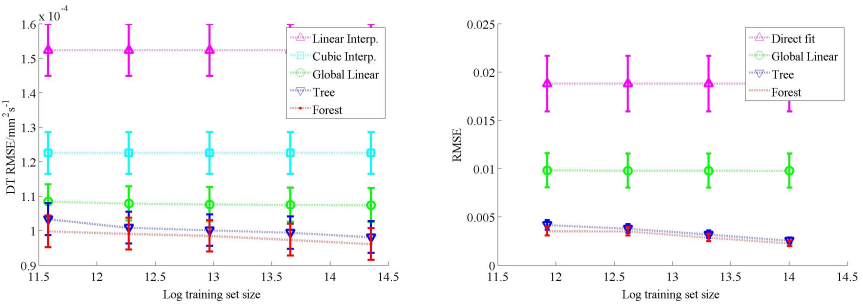


Fig. 3. Left: reconstruction errors against ground truth for various reconstructed DT maps, after downsampling by a factor of 2 and reconstructing with $m = 2$, as a function of training set size. Right: reconstruction errors for NODDI parameter reconstructions. In both cases, $n = 2$, and the error bars show standard deviation over subjects.

Figure 4 shows qualitative results of both image quality transfer applications using mappings learned from HCP data to enhance a non-HCP data set. The data set comes from a standard 3T clinical scanner. It has a single HARDI shell with $b = 1000\,\mathrm{s\,mm^{-2}}$ and 30 gradient directions. The voxel size is $2^3\,\mathrm{mm}^3$. Image quality transfer sharpens weak structures in super-resolution and produces plausible NODDI parameter maps.

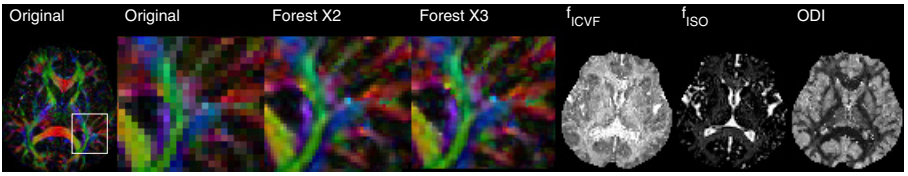


Fig. 4. Image quality transfer to a non-HCP data set. From left to right: (i) Colour FA map of original data; (ii) zoomed in view of boxed area; (iii) and (iv) forest up sampling with $m = 2$ and $m = 3$; (v)-(vii) reconstructed NODDI parameter maps.

4 Discussion

The super-resolution results show that patches in low resolution diffusion images contain a significant amount of information about the subvoxel content at the centre of the patch. Thus, with a transformation of sufficient complexity, we can predict high resolution images with much greater accuracy than standard interpolation techniques. Further work is required to make a formal performance comparison with other super-resolution techniques, in particular [5], but a key novelty of our implementation is to use DT patches as input rather than DWI patches and to output patches of parameter estimates rather than raw DWIs. An advantage of outputting fitted model patches is to constrain the output image structure to realistic local configurations. Moreover, treating each DWI separately, as in [5], fails to exploit the strong correlation among DWIs, which is highly informative. For example, enhancing the raw DWIs and then fitting the NODDI model fails because it still fits to only a single b; it is the neighbourhood structure of the fitted DTs that informs on the parameters. One might consider using patches of the full collection of raw DWIs as input or output. However, this has two practical problems: i) it complicates transfer to sparser data sets where the set of input DWIs is, by definition, different to that of the high quality training data; ii) the memory requirements for training become orders of magnitude larger, as, in the super-resolution example, p_1 increases from 6 to 288. However, input patches of parameters of more complex models than the DT, such as higher-order spherical harmonic coefficients, may improve performance.

Random-forest regression shows a dramatic improvement over both standard parameter estimation and global linear regression in recovering NODDI parameter maps from single-shell HARDI data sets and produces plausible maps of all three important parameters.

Training times can be considerable for regression trees (around 1 day of processor time for the largest training sets in figure 3 with unoptimised matlab code), but reconstruction times are small (a few minutes for a full volume), as they require just a linear transformation in each voxel. This is much quicker in fact than direct NODDI fitting, which is non-linear and requires several hours of processor time per image volume. The method extends naturally to predict parameters of other models, such muti-fibre models [9], where computational advantages again are potentially significant. More broadly, the framework extends naturally beyond diffusion imaging to any scalar or vector valued images.

Imperfections remain in reconstructed images in both applications. Various refinements of the random forest regression may improve performance. Reconstruction error reduces rapidly as the number of trees increases from 1 to 3, but stabilises above 4 so that the 8 trees we use here is sufficient. For fixed training set size, little improvement arises from using more than 2 source images, so our collection of 8 seems sufficient. Performance increases with n for a fixed number of data points in the training set, suggesting that neighbourhoods are informative even several steps from the output voxel. However, the data points are larger at higher n, so memory limits occur at fewer data points; $n = 2$ is a good compromise with the current implementation. Choice of features affects perfor-

mance. From our set, features based on the largest and smallest DT eigenvalues dominate near the tree root; the second eigenvalue rarely appears. Orientational features are important for finer partitions nearer leaf nodes. Smaller n leads to more complex trees, because each linear transformation has less parameters. Other high dimensional regression techniques may improve results. Moreover, constraining the mapping with a data fitting term would be beneficial, but leads to non-linear reconstruction increasing computation times.

In summary, this initial formulation and demonstration of image quality transfer shows compelling results from a simple implementation that improves significantly on standard interpolation and estimation techniques. Further work must establish sufficiency of reliability to make downstream improvements in practical applications such as tractography and image-based biomarker studies. Both random-forest regression and the local patch-based image representation lend themselves well to generalizability in the presence of pathology or other effects not represented in the training set at least for diffuse or macroscopic effects. However, further work needs to evaluate performance in such situations.

Acknowledgements. EPSRC grants E007748 and I027084 supported this work. Data were provided [in part] by the HCP, WU-Minn Consortium (PIs: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by NIH and Wash. U.

References

1. Sotiropoulos, S.N., et al.: Advances in diffusion MRI acquisition and processing in the human connectome project. NeuroImage 80, 125–143 (2013)
2. Zhang, H., et al.: NODDI: Practical in vivo neurite orientation dispersion and density imaging of the human brain. NeuroImage 61, 1000–1016 (2012)
3. Rousseau, F.: Brain Hallucination. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 497–508. Springer, Heidelberg (2008)
4. Rueda, A., Malpica, N., Romero, E.: Single-image super-resolution of brain MR images using overcomplete dictionaries. Med. Im. An. 17(1), 113–132 (2013)
5. Coupé, P., Manjón, J.V., Chamberland, M., Descoteaux, M., Hiba, B.: Collaborative patch-based super-resolution for diffusion-weighted images. NeuroImage 83, 245–261 (2013)
6. Ye, D.H., Zikic, D., Glocker, B., Criminisi, A., Konukoglu, E.: Modality propagation: coherent synthesis of subject-specific scans with data-driven regularisation. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part I. LNCS, vol. 8149, pp. 606–613. Springer, Heidelberg (2013)
7. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
8. Criminisi, A., Shotton, J.: Decision forests for computer vision and medical image analysis. Springer (2013)
9. Seunarine, K.K., Alexander, D.C.: Multiple fibers: beyond the diffusion tensor. In: Johansen-Berg, H., Behrens, T.E.J. (eds.) Diffusion MRI: from Quantitative Measurement to in Vivo Neuroanatomy, pp. 56–74. Academic Press (2009)

Complete Set of Invariants of a 4^{th} Order Tensor: The 12 Tasks of HARDI from Ternary Quartics

Théo Papadopoulo, Aurobrata Ghosh, and Rachid Deriche

Athena Project Team, Inria Sophia Antipolis - Méditerranée, France*

Abstract. Invariants play a crucial role in Diffusion MRI. In DTI (2^{nd} order tensors), invariant scalars (FA, MD) have been successfully used in clinical applications. But DTI has limitations and HARDI models (e.g. 4^{th} order tensors) have been proposed instead. These, however, lack invariant features and computing them systematically is challenging.

We present a simple and systematic method to compute a *functionally complete set* of invariants of a non-negative 3D 4^{th} order tensor with respect to SO_3. Intuitively, this transforms the tensor's non-unique ternary quartic (TQ) decomposition (from Hilbert's theorem) to a unique canonical representation independent of orientation – the invariants.

The method consists of two steps. In the first, we reduce the 18 degrees-of-freedom (DOF) of a TQ representation by 3-DOFs via an orthogonal transformation. This transformation is designed to enhance a rotation-invariant property of choice of the 3D 4^{th} order tensor. In the second, we further reduce 3-DOFs via a 3D rotation transformation of coordinates to arrive at a canonical set of invariants to SO_3 of the tensor.

The resulting invariants are, by construction, (i) functionally *complete*, (ii) functionally *irreducible* (if desired), (iii) computationally *efficient* and (iv) *reversible* (mappable to the TQ coefficients or shape); which is the novelty of our contribution in comparison to prior work.

Results from synthetic and real data experiments validate the method and indicate its importance.

Keywords: Invariants, SO_3, 4^{th} order tensors, ternary quartics, orthogonal & rotation transforms, canonical representation.

1 Introduction

High angular resolution diffusion imaging (HARDI) has vastly improved our analysis of the brain's microstructure and the detection of crossing-fibers from diffusion MRI (dMRI) where classical diffusion tensor imaging (DTI) is limited. Nonetheless, to assess the integrity of the white-matter affected by development, aging or neuro-degenerative pathologies, it is crucial to compute rotation invariant scalars or biomarkers. Although numerous invariant scalars are known for the 2^{nd} order DTI tensor $\mathbf{D}$, e.g. FA, MD, etc. [1] and have been successfully

* This work has been partially supported by the ANR project Mosifah (MN Program).

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 233–240, 2014.

used in clinical applications, few invariant scalars are available for HARDI. Although richer in description, higher order HARDI models lack comprehensive invariant scalar descriptors to decipher that information. Therefore, computing HARDI invariants is an important problem.

HARDI models are often estimated in the spherical harmonic (SH) or the bijective higher order tensor (HOT) bases. (HOTs are homogeneous polynomials and their SH transform provides a bijective map to SHs). Thus, studying invariant features of the SH or HOT bases has wide applications. Early proposals of HARDI biomarkers based on these bases can be found in [2, 3]. Though these are popular, they only recover a few of the possible invariant scalar biomarkers.

In the case of the 4^{th} order tensor (HOT4), equivalently SHs, more systematic approaches are presented in [4–6] but only six of the twelve invariants are found. [4] also presents D-eigenvalues which are the extrema of the HOT4. Further, [6] introduces the *integrity basis* and the idea of polynomial completeness.

From this, a more general idea presents itself – ***functional completeness:*** "find a set of invariants such that all other invariants are functions of the invariants of the set." Naturally, this leads to the idea of ***functional irreducibility:*** "what is the smallest such set?" How to find *the minimal set of invariants* required to completely describe the shape of a HOT4? The importance of this question can be illustrated on the DTI tensor. A functionally complete & irreducible invariant set of $\mathbf{D}$ is its eigenvalue-set $\{\lambda_1, \lambda_2, \lambda_3\}$. However, if only $\{\lambda_1, \lambda_2\}$ are found and any number of invariant functions of $\{\lambda_1, \lambda_2\}$ (even greater than 3), it is impossible to describe the shape of $\mathbf{D}$ without λ_3. And there exist infinitely many such invariant functions!

Further noteworthy results for HOT4 are proposed in [7], where the cardinality of the irreducible set is shown to be 12. The paper presents polynomial invariants and tries to establish polynomial completeness. However, this is only partially successful and the polynomial formulae are intractable and cannot be inverted to recompute the HOT4 coefficients, hence the approach is ***irreversible***. [8] is based on rank-4 SHs and proposes a set of 25 invariants (and ad-hoc functions of these). However, no proof of functional completeness, irreducibility or reversibility is attempted (since only 12 invariants are required).

In this paper, we consider non-negative HOT4s and their Ternary Quartic (TQ) parameterization from Hilbert's theorem [9, 10] since the physical constraint of positivity arises commonly in dMRI, e.g. the diffusion kurtosis tensor (DKT) [11], the apparent diffusion coefficient tensor (ADC), or the Cartesian tensor fiber orientation distribution (CT-FOD).

We propose a simple and systematic method to compute a *functionally complete* set of invariants to SO_3 of a non-negative HOT4 by mapping its non-unique TQ decomposition to a unique canonical representation independent of orientation. From this we extract the invariants. This involves a two step reduction process via an orthogonal and a rotation transform. The resulting invariants are, by construction, (i) ***functionally complete***, (ii) ***functionally irreducible*** (if desired), (iii) *efficient* to compute and (iv) ***reversible***. Reversibility ensures that the shape of the TQ can be inferred from the invariants and all prior/other

invariants can be expressed in terms of these new ones. We validate the method on synthetic and real data tests to highlight its importance.

2 Materials and Methods

Background. Positivity is a common constraint in dMRI. The right way to parameterize a non-negative 3D 4^{th} order tensor $\mathcal{A}$, is a sum-of-squares as shown by Hilbert's theorem on non-negative TQ's [9, 10]:

$$F_4(\mathbf{x}) = \sum_{i+j+k=4} \mathcal{A}_{i,j,k} x^i y^j z^k = (\mathbf{v}^T \mathbf{w}_1)^2 + (\mathbf{v}^T \mathbf{w}_2)^2 + (\mathbf{v}^T \mathbf{w}_3)^2 = \mathbf{v}^T \mathbf{W}\mathbf{W}^T \mathbf{v},$$

where $\mathbf{x}^T = [x, y, z]$, $\mathbf{v}^T = [x^2, y^2, z^2, \sqrt{2}xy, \sqrt{2}xz, \sqrt{2}yz]$, $\mathbf{w}_i$ $(i = 1..3)$ are 6×1 vectors of the coefficients of the three quadratic forms and $\mathbf{W} = [\mathbf{w_1}|\mathbf{w_2}|\mathbf{w_3}]$ is a 6×3 matrix. The TQ parameterization has 18=6×3 independent coefficients, although $\mathcal{A}$ has only 15. These 3 degrees-of-freedom (DOFs) imply that $\mathbf{W}$ and $\mathbf{WS}$ for any 3×3 orthogonal matrix $\mathbf{S}$ (3-DOFs) result in the same F_4. To resolve this ambiguity, [9] proposed a QR/Iwasawa decomposition of $\mathbf{A}$, the top 3×3 block of $\mathbf{W}$, which fixes $\mathbf{S}$ such that $\mathbf{A}$ becomes triangular – effectively zeroing out 3 coefficients. Here, we introduce a slightly different notation:

$$F_4(\mathbf{x}) = \sum \left(\mathbf{x}^T \mathbf{C}_i \mathbf{x}\right)^2 = ||\mathbf{c}||^2, \quad i = 1..3, \tag{1}$$

where $\mathbf{C}_i$ are 3×3 symmetric matrix representations of vectors $\mathbf{w}_i$ and $\mathbf{c}^T = [\mathbf{x}^T\mathbf{C}_1\mathbf{x}, \mathbf{x}^T\mathbf{C}_2\mathbf{x}, \mathbf{x}^T\mathbf{C}_3\mathbf{x}]$. In this notation $F_4(\mathbf{x}) = \mathbf{c}^T\mathbf{c} = \widehat{\mathbf{c}}^T\widehat{\mathbf{c}}$, where $\widehat{\mathbf{c}} = \mathbf{S}\mathbf{c}$.

Theory. TQ's have 18-DOFs, while HOT4s have 15-DOFs. The excess 3-DOFs can be eliminated by fixing an orthogonal transformation $\mathbf{S}$. Additionally a rotation invariance criterion would further eliminate 3-DOFs via a rotation transformation $\mathbf{R}$, of coordinates, resulting in the known 12-DOFs.

The basic idea is, therefore, a two step reduction process to twice remove 3-DOFs – first an orthogonal transform $\mathbf{S}$ and second a rotation transform $\mathbf{R}$ – resulting in a canonical representation of a TQ invariant to rotations: $\underbrace{\mathbf{C}_{18\text{-DOF}} \rightsquigarrow \mathbf{S}_{3\text{-DOF}}}_{\text{Step-1}} \Longrightarrow \underbrace{\mathbf{C'}_{15\text{-DOF}} \rightsquigarrow \mathbf{R}_{3\text{-DOF}}}_{\text{Step-2}} \Longrightarrow \mathbf{C''}_{12\text{-DOF}} \longmapsto$ 15 invariants (complete set) $\dashrightarrow$ 12 invariants (complete & irreducible set if desired).

Orthogonal Transform ($\mathbf{S}$). Contrary to [9], we do not choose $\mathbf{S}$ to zero out 3 coefficients, since that is not invariant to SO_3 and the zeros would be lost after the second rotation transform step. Instead, we design $\mathbf{S}$ to enhance a property of choice that is rotation-invariant. First we provide an overview of the approach before describing the rotation-invariant properties.

The application of $\mathbf{S}$ on $\mathbf{c}$ results in the transformations $\mathbf{C'}_i = \sum_j S_{ij}\mathbf{C}_j, i, j = 1..3$. Hence, each $\mathbf{C'}_i(\sigma_i)$ is a function of the i^{th} row vector $\sigma_i = [S_{i1}, S_{i2}, S_{i3}]$ of $\mathbf{S}$. If $\mathbf{S}$ is required to enhance a chosen property $\mathcal{P}$ of $\{\mathbf{C'}_i\}_{i=1}^3$ and if $\mathcal{P}$ is a quadratic function, then $\mathcal{P}$ can be written as: $\mathcal{P}(\{\mathbf{C'}_i(\sigma_i)\}) = \frac{1}{2}\sigma\mathbf{K}\sigma^T$,

where $\mathbf{K}$ is the Hessian of the quadratic function $\mathcal{P}$. To enhance $\mathcal{P}$ it suffices to find its extrema or to diagonalize $\mathbf{K} = \mathbf{U}\Lambda\mathbf{U}^T$. $\mathbf{K}$ being symmetric, $\mathbf{U}$ is an orthogonal matrix whose columns either maximize or minimize $\mathcal{P}$ depending on the order of the corresponding eigenvalues in Λ. Hence, if $\mathbf{S}$ is chosen to be $\mathbf{U}^T$, $\mathbf{S}$ would enhance $\mathcal{P}$ when computing $\{\mathbf{C'}_i\}$ from $\{\mathbf{C}_i\}$. In summary, to enhance a quadratic property via an orthogonal transform $\mathbf{S}$, we simply need to choose $\mathbf{S}$ as the eigenvector matrix $\mathbf{U}^T$ from the eigen-decomposition of $\mathbf{K}$.

Let us now examine rotation invariant properties of $\{\mathbf{C'}_i\}$. Since, $\{\mathbf{C'}_i\}$ are all 3×3 symmetric matrices, their rotation-invariant properties are simply the coefficients of their characteristic polynomials: $M_1(\mathbf{C'}_i(\sigma_i)) = trace(\mathbf{C'}_i(\sigma_i))$, $M_2(\mathbf{C'}_i(\sigma_i)) = \sum_{k_1,k_2} \mu_{k_1}\mu_{k_2}$, where μ_k are the eigenvalues of $\mathbf{C'}_i(\sigma_i)$, and $M_3(\mathbf{C'}_i(\sigma_i)) = det(\mathbf{C'}_i(\sigma_i))$. M_1 is a linear function of σ_i, M_2 is a quadratic function, while M_3 a cubic function. Hence M_1^2 & M_2 are two quadratic function-properties (i.e. $\mathcal{P}$) that are also rotation-invariant.

Let us name $\mathbf{H}_1$ and $\mathbf{H}_2$ the two Hessian matrices of M_1^2 and M_2 respectively. To compute the corresponding orthogonal transforms $\mathbf{S}$ that enhance these properties we need to calculate $\mathbf{H}_1$ and $\mathbf{H}_2$. $\mathbf{H}_1$ is simply $\mathbf{H}_1 = 2\mathbf{T}\mathbf{T}^T$, where $\mathbf{T}^T = [trace(\mathbf{C'}_1), trace(\mathbf{C'}_2), trace(\mathbf{C'}_3)]$. The form of $\mathbf{H}_2$ is more involved but is simple to derive on a computer-algebra-system and can be provided on request.

Interestingly, any combination of M_1 and M_2 that is a quadratic function is also a rotation invariant property. Linear combinations of M_1^2 and M_2 are simple to compute as linear combinations of $\mathbf{H}_1$ and $\mathbf{H}_2$. In particular, we consider the following invariant properties:

- **Trace of $\mathbf{C}'_i$:**
  $\mathcal{P}_1(\sigma) = trace(\mathbf{C}'_i)^2 = M_1(\mathbf{C}'_i)^2 = \frac{1}{2}\sigma\mathbf{H}_1\sigma^T = \frac{1}{2}\sigma\mathbf{K}_1\sigma^T$.
- **Variance of the eigenvalues of $\mathbf{C}'_i$:**
  $\mathcal{P}_2(\sigma) = M_1(\mathbf{C}'_i)^2 - 3M_2(\mathbf{C}'_i) = \frac{1}{2}\sigma\left(\mathbf{H}_1 - 3\mathbf{H}_2\right)\sigma^T = \frac{1}{2}\sigma\mathbf{K}_2\sigma^T$.
- **Frobenius norm of $\mathbf{C}'_i$:**
  $\mathcal{P}_3(\sigma) = M_1(\mathbf{C}'_i)^2 - 2M_2(\mathbf{C}'_i) = \frac{1}{2}\sigma\left(\mathbf{H}_1 - 2\mathbf{H}_2\right)\sigma^T = \frac{1}{2}\sigma\mathbf{K}_3\sigma^T$.

Thus, it is possible to enhance any of these properties while designing an orthogonal transform $\mathbf{S}$ in the first reduction step. However, this is a generic design and many other properties could be considered. $\mathcal{P}_2$ in particular is the same as the numerator of FA for DTI. However, in this case $\{\mathbf{C}'_i\}$ may have negative eigenvalues since their squares are considered in the TQ representation. Here, we only present results of $\mathbf{K}_2$ though we have experimented with all three properties.

It is important to note that in choosing $\mathbf{S}$ to enhance a rotation-invariant property, contrary to [9], the application of $\mathbf{S}$ to the TQ will not result in 3 coefficients becoming null. It will reduce 3-DOFs of the TQ but will still result in 18 non-null coefficients, where 3 are now dependent.

Rotation Transform ($\mathbf{R}$). The second reduction step involves further eliminating 3-DOFs via a rotation transformation of the coordinates to map the modified TQ coefficients $\{\mathbf{C}'_i\}$ to a representation that is independent of orientation. We choose this rotation transform $\mathbf{R}$ from the eigen-decomposition of $\mathbf{C}'_1 = \mathbf{R}\Sigma\mathbf{R}^T$.

$\mathbf{C}'_1$ is diagonalized and the other two $\mathbf{C}'_i$'s are transformed as $\mathbf{C}''_i = \mathbf{R}^T\mathbf{C}'_i\mathbf{R}$. The two transformations, $\mathbf{S}$ and $\mathbf{R}$ combined, map the original TQ parameterization to a canonical representation that is invariant to all 3D rotations.

This results in 15 non-null coefficients ($\mathbf{C}''_1 = \Sigma : 3, \mathbf{C}''_{2\&3} : 2 \times 6$) with 12-DOFs, implying that three are dependent. This, however, ensures functional completeness and, by construction, the invariants are also reversible and still maintain the TQ structure.

The process for extracting the irreducible set of 12 invariants involves considering the off-diagonal terms of the matrix $\mathbf{K}$. This yields a linear system in the coefficients of $\mathbf{C}''_i$, which can be solved to express 3 coefficients as an expression of the others. However, the choice of which 3 coefficients to eliminate is not unique. This ambiguity implies that the space of TQs is covered by several maps and that there are many ways to extract the irreducible set of 12 invariants from the 15 non-null (invariant) coefficients. Therefore, although for theoretical purposes it is possible to extract the irreducible set of 12 invariants, in practice it is more convenient to work directly with the 15 invariant coefficients with slightly redundant information but guaranteed functional completeness.

Unicity of Sign. The two steps involving $\mathbf{S}$ and $\mathbf{R}$ proceed via eigen-decompositions of $\mathbf{K}$ and $\mathbf{C}'_1$, implying that the eigenvectors in $\mathbf{U}$ and $\mathbf{R}$ are unique only up to a sign. This introduces an ambiguity since a large set of maps can be used to transform the TQ coefficients to the canonical representation. These arbitrary sign changes can become difficult to handle in the presence of noise or when values are close to zero. We devised the following procedure to minimize this ambiguity. The sign of the columns of $\mathbf{U}$ are flipped to ensure that the largest absolute eigenvalue of the corresponding $\mathbf{C}'_i$ is positive. The sign of the columns of $\mathbf{R}$ are flipped to ensure that the largest absolute entry in the column is positive. However, since $\mathbf{R}$ is a rotation matrix, if the above operation results in $det(\mathbf{R}) < 0$, the sign of $\mathbf{R}$ is flipped to guarantee a proper rotation.

So far in this section, we presented the coefficients of the quadratic forms (TQ coefficients) modified by the orthogonal and rotation transforms as the invariants. However, any functions of these invariant coefficients are also valid invariants. In practice, on real data, we found that working with "super-features" such as M_1, M_2, M_3 of the final canonical coefficients to be more robust to rotation tests. Note that these still satisfy functional completeness but reversibility becomes harder to establish. Nonetheless, it is important to explore suitable "super-feature" functions, especially from a physical/physiological perspective.

3 Experiments and Results

We conducted tests on synthetic and real data to verify the invariance of the canonical invariants to rotation. In the synthetic data experiments, we used the multi-tensor model to generate voxels with various crossing configurations (1, 2 & 3). These were then arbitrarily rotated in space before estimating the HOT4s and the invariants. We conducted a similar real data experiment from an in vivo acquisition where we rotated the set of gradient directions arbitrarily before

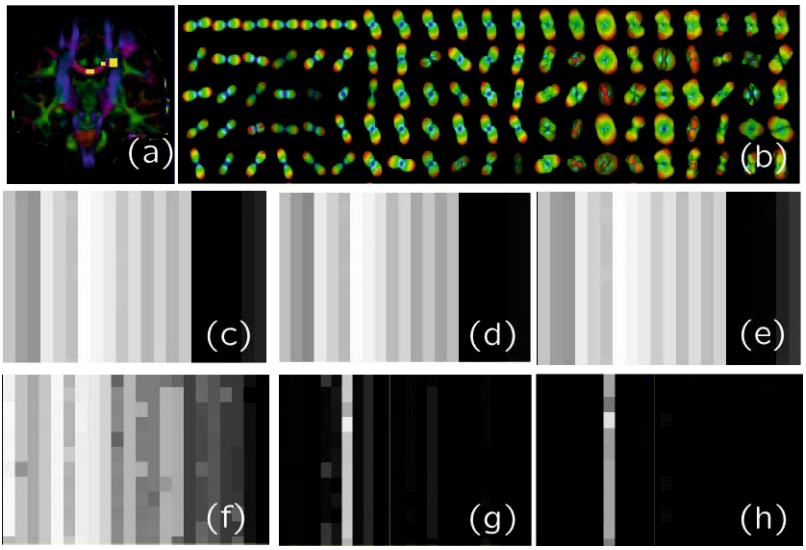


Fig. 1. Real data (qualitative): Rotation invariance test and comparison with principal invariants (PIs). (a) Coronal slice indicating chosen voxels. (b) 1st row: the sample voxels, followed by samples with random rotations. (c-e): 3 PIs. (f-h) 3 new/canonical invariants. PIs (c-e) resemble each other or capture similar information. The new invariants (f-h) have a richer spectrum and as expected capture more information.

estimating the HOT4s and the invariants. In both cases, the results were very similar. Hence, we present only the real data results here.

The real data was acquired with a whole-body 3 Tesla Magnetom TRIO scanner (Siemens Medical Solution). It was equipped with an 8-channel head array coil. The twice-refocused spin-echo EPI sequence (TR = 12s, TE = 100 ms, 128×128 image matrix, FOV = 22.0 = 22.0 cm^2) consisted of 60 diffusion-encoding gradient directions with a b-value of 1000s/mm^2 [12].

In the experiment, we chose 21 voxels from a coronal slice with some voxels in the Corpus Callosum (CC) and others in a region where radial projections of the CC intersect the superior longitudinal fasciculus (SLF) and the cortico-spinal tract (CST), Fig. 1a. This allowed us to consider voxels with 1, 2 & 3 fiber crossings. For each of these voxels, we randomly rotated the acquisition gradients 50 times to generate 50 test cases. From these, we estimated the non-negative HOT4s or TQs, Fig. 1b, and computed the principal invariants [6] and the new/canonical invariants. Some of these are in Figs. 1c-h. Finally, we computed the difference between each pair of 50 test cases for all the voxels to compute the average relative error for both the principal & new invariants. This quantitative result is presented in Fig. 2.

In Figs. 1c-h we notice that the principal invariants (mid-row) capture almost the same information, while the new invariants (bot-row) have a richer spectrum. The new invariants look more noisy but this is only due to the visual

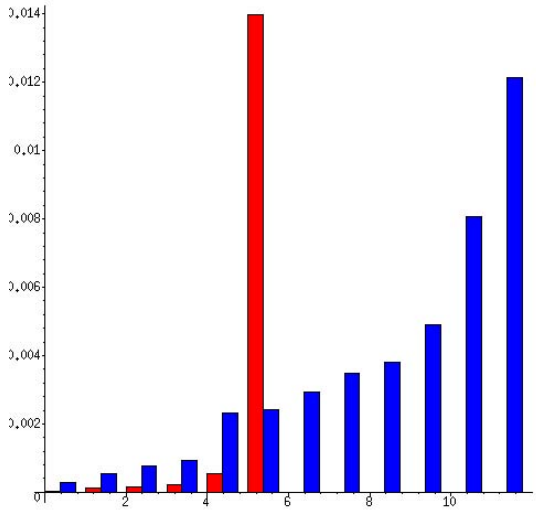


Fig. 2. Real data (quantitative): Rotation invariance test and comparison with principal invariants (PIs). Average relative errors between the new/canonical invariants compared to PIs (see Fig. 1). Blue-bars: new/canonical invariants (only a few are shown). Red-bars: principal invariants. All invariants (new & PIs) are stable under rotation and commit less than 2% error even in the worst case.

representation. From Fig. 2 it is clear that all the (new & principal) invariants are stable under rotation and commit less than 2% error on an average even in the worst case. Thus, our tests reveal that the new invariants are stable and as expected capture more invariant/shape information than principal invariants.

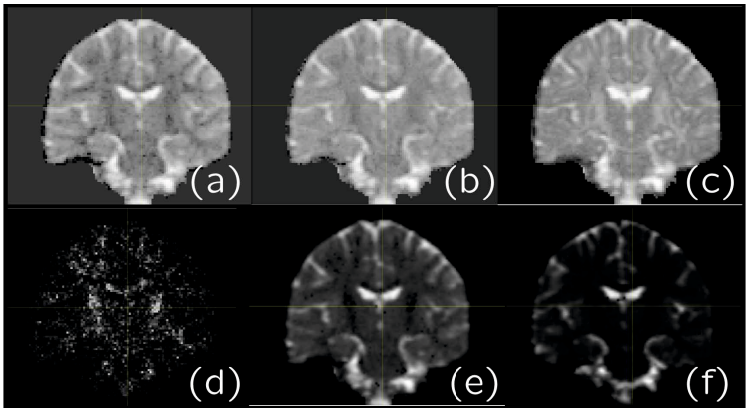


Fig. 3. A sample of the new/canonical invariants. Six are displayed on a coronal slice.

4 Conclusion

We have proposed a simple and systematic method for computing a functionally complete set of invariants of a non-negative HOT4. The method has two steps. First, we designed an orthogonal transform to enhance rotation-invariant properties of the tensor while eliminating 3-DOFs from its TQ representation. Second, we chose a rotation transform to map the TQ coefficients to a canonical representation invariant to orientation. From this, we extracted the invariants.

These canonical invariants are, by construction, (i) ***functionally complete***, (ii) ***functionally irreducible*** (if desired) – although in practice we deal with a slightly redundant super-set, (iii) efficient to compute and (iv) ***reversible*** – allowing us to map the TQ's shape completely. Reversibility ensures that all other/prior invariants can be described as functions of these canonical invariants.

We conducted tests on synthetic and real data and validated the invariants. The results revealed that the new invariants were stable under rotation and captured a rich spectrum of information. In the future, we plan to explore physiological interpretations and to conduct a more elaborate clinical validation.

References

1. Basser, P.J.: Inferring microstructural features and the physiological state of tissues from diffusion-weighted images. NMR in Biomedicine 8, 333–344 (1995)
2. Özarslan, E., Vemuri, B.C., Mareci, T.H.: Generalized scalar measures for diffusion MRI using trace, variance and entropy. Mag. Res. in Med. 53(4), 866–876 (2005)
3. Tuch, D.: Q-ball imaging. Mag. Res. in Med. 52(6), 1358–1372 (2004)
4. Qi, L., Han, D., Wu, E.: Principal invariants and inherent parameters of diffusion kurtosis tensors. J. of Math.l Anal. and Ap. 349(1), 165–180 (2009)
5. Fuster, A., van de Sande, J., Astola, L., Poupon, C., Velterop, J., Romeny, B.M.t.H.: Fourth-order Tensor Invariants in High Angular Resolution Diffusion Imaging. In: Zhang, G., Adluru, N (eds.) CDMRI Workshop, MICCAI. LNCS, vol. 6891. Springer (2011)
6. Ghosh, A., Papadopoulo, T., Deriche, R.: Biomarkers for HARDI: 2nd & 4th order tensor invariants. ISBI, Barcelona (May 2012)
7. Ghosh, A., Papadopoulo, T., Deriche, R.: Generalized invariants of a 4th order tensor: Building blocks for new biomarkers in dMRI. In: Panagiotaki, E., ODonnell, L., Schultz, T., Zhang, G.H. (eds.) CDMRI Workshop, MICCAI, pp. 165–173 (2012)
8. Schwab, E., Çetingül, H.E., Afsari, B., Yassa, M.A., Vidal, R.: Rotation invariant features for HARDI. In: Gee, J.C., Joshi, S., Pohl, K.M., Wells, W.M., Zöllei, L. (eds.) IPMI 2013. LNCS, vol. 7917, pp. 705–717. Springer, Heidelberg (2013)
9. Barmpoutis, A., Hwang, M.S., Howland, D., Forder, J.R., Vemuri, B.C.: Regularized positive-definite fourth-order tensor field estimation from DW-MRI. NeuroImage 45(1), S153–S162 (2009)
10. Ghosh, A., Moakher, M., Deriche, R.: Ternary quartic approach for positive 4th order diffusion tensors revisited. ISBI: From Nano to Macro, pp. 618–621 (June 2009)
11. Barmpoutis, A., Zhuo, J.: Diffusion kurtosis imaging: Robust estimation from DW-MRI using homogeneous polynomials. In: ISBI 2011, pp. 262–265 (2011)
12. Anwander, A., Tittgemeyer, M., von Cramon, D.Y., Friederici, A.D., Knosche, T.R.: Connectivity-Based Parcellation of Broca's Area. Cerebral Cortex 17(4), 816–825 (2007)

In vivo Estimation of Dispersion Anisotropy of Neurites Using Diffusion MRI

Maira Tariq[1], Torben Schneider[2], Daniel C. Alexander[1], Claudia A.M. Wheeler-Kingshott[2], and Hui Zhang[1]

[1] Centre for Medical Image Computing and Department of Computer Science, University College London, United Kingdom
[2] NMR Research Unit, Department of Neuroinflammation, Institute of Neurology, University College London, United Kingdom

Abstract. We present a technique for mapping dispersion anisotropy of neurites in the human brain *in vivo*. Neurites are the structural substrate of the brain that support its function. Measures of their morphology from histology provide the gold standard for diagnosing various brain disorders. Some of these measures, e.g. neurite density and orientation dispersion, can now be mapped *in vivo* using diffusion MRI, enabling their use in clinical applications. However, *in vivo* methods for estimating more sophisticated measures, such as dispersion anisotropy, have yet to be demonstrated. Dispersion anisotropy allows more refined characterisation of the complex neurite configurations such as fanning or bending axons; its quantification *in vivo* can offer new imaging markers. The aim of this work is to develop a method to estimate dispersion anisotropy *in vivo*. Our approach builds on the Neurite Orientation Dispersion and Density Imaging (NODDI), an existing clinically feasible diffusion MRI technique. The estimation of dispersion anisotropy is achieved by incorporating Bingham distribution as the neurite orientation distribution function, with no additional acquisition requirements. We show the first *in vivo* maps of dispersion anisotropy and demonstrate that it can be estimated accurately with a clinically feasible protocol. We additionally show that the original NODDI is robust to the effects of dispersion anisotropy, when the the new parameter is not of interest.

1 Introduction

Axons and dendrites, collectively known as neurites, are the projections from the cell body of a neuron; they are the structural underpinnings of brain functions. The morphology of neurites, quantified using histological analysis of postmortem tissue, provides the gold standard for understanding the development [1], function [2] and pathology [3] of the brain. Accessing such information non-invasively has been of great interest because this will enable a dynamic view of the brain in health and disease. Diffusion MRI (dMRI) is such a technique, which can probe the microstructure and is becoming an indispensable tool for assessing the structure of neurites, *in vivo*.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 241–248, 2014.

The standard dMRI technique, diffusion tensor imaging (DTI) [4], provides sensitivity to neurite morphology, but can not quantify specific measures such as neurite density and orientation dispersion. Jespersen et al. proposed the first dMRI technique to estimate these neurite measures directly [5] in *ex vivo* imaging; their technique was subsequently validated with detailed histological comparison [6]. Zhang et al enabled the *in vivo* mapping of these measures with the development of the neurite orientation dispersion and density imaging (NODDI) [7]. The clinical feasibility of NODDI has allowed the quantification of neurite morphology to gain adoption in neuroimaging research, e.g. to study epilepsy [8] and brain development [9].

However, one limitation of NODDI is that it can not characterise complex neurite configurations such as those arising from fanning and bending axons. NODDI models orientation distribution of neurites with the Watson distribution [10], which constrains the dispersion about the dominant orientation to be isotropic (see Fig.1). For fanning and bending axons, the dispersion about the dominant orientation is typically anisotropic: the dispersion is the highest along the plane of fanning and bending but the lowest perpendicular to the plane. *Ex vivo* imaging has shown such dispersion anisotropy is widespread in the brain [11] but can not be accessed with NODDI or other *in vivo* techniques. Quantifying this anisotropy *in vivo* will provide a more refined measure of neurite morphology that can serve as a potential imaging marker of neurite integrity [12] and be used to improve tractography [11,13].

The aim of this work is to develop a method that can estimate dispersion anisotropy of neurites *in vivo*.We take the approach of building on NODDI because it provides a dMRI acquisition protocol that has been demonstrated to be clinically feasible [8,9]. We propose a new NODDI model that incorporates the Bingham distribution [10] to enable the quantification of dispersion anisotropy of neurites. Bingham distribution has been used in various dMRI techniques [14,11], but the focus has been on mapping brain connectivity. Our aim is to map microstructure using biophysically meaningful parameters. The key advance in this work over these approaches is the use of multi-shell data, as in [7], which enables estimation of microstructure at the same time as the fibre dispersion parameters. Our proposed model enables estimation of dispersion anisotropy and the primary dispersion orientation, along with the estimates of

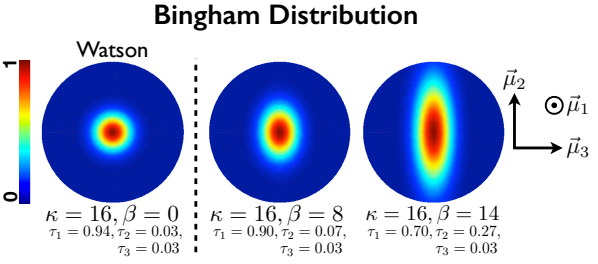


Fig. 1. Probability density plots for Bingham distribution. From left to right, increasing dispersion anisotropy about the dominant orientation $\boldsymbol{\mu}_1$. The primary dispersion orientation, $\boldsymbol{\mu}_2$, represents the orientation of dispersion anisotropy about $\boldsymbol{\mu}_1$. Watson is a special case of Bingham distribution.

neurite density and their concentration along the dominant orientation, without imposing any additional acquisition requirements compared to the original model. We define a new measure of dispersion anisotropy based on the *Orientation Tensor* (OT), following the work by [12]. To assess the performance of the proposed NODDI model, we evaluate the accuracy and precision of estimating dispersion anisotropy, with the 2-shell optimised protocol [7]. We also assess the consequences of not accounting for dispersion anisotropy, on estimation of the NODDI parameters.

2 NODDI Tissue Model

NODDI relates the microstructure parameters to acquired dMRI signals, using a two-level multi-compartment model. At the first level, the signals from the tissue and non-tissue components of the brain are modelled separately, weighted by their respective volume fractions, to obtain the total signal.

The non-tissue compartment represents the free diffusing water in the brain (e.g. CSF and interstitial fluid) and is modelled by isotropic Gaussian diffusion, with diffusivity, $d_{iso} = 3.0 \times 10^{-9} m^2 s^{-1}$. The volume fraction of this compartment is denoted by ν_{iso} and that of the tissue compartment by $(1 - \nu_{iso})$.

The second level models the signal from tissue compartment (grey and white matter (GM/WM)) as a sum of the signal originating from inside the neurites (intra-neurite) and that from the space outside them (extra-neurite), weighted by their respective volume fractions. The intra-neurite volume fraction gives an estimate of neurite density and we denote it by ν_{in}, while the extra-neurite volume fraction is $(1-\nu_{in})$. The membrane of neurites restrict the water diffusion to be along their length, so the signal from a neurite is computed as the attenuation due to unhindered diffusion along a stick with diffusivity $d_{\parallel} = 1.7 \times 10^{-9} m^2 s^{-1}$. The intra-neurite signal, A_{in} is:

$$A_{in} = \int_{\mathbb{S}^2} f(\boldsymbol{n}) e^{-b d_{\parallel} (\boldsymbol{q}.\boldsymbol{n})^2} \, \mathrm{d}\boldsymbol{n}, \tag{1}$$

where $\boldsymbol{q}$ and b are the gradient direction and b-value of the diffusion-weighting, and $f(\boldsymbol{n})d\boldsymbol{n}$ gives the probability of finding sticks along an orientation $\boldsymbol{n}$. The diffusion in the extra-neurite space is hindered by the presence of neurites; thus the orientation distribution of neurites affects the the extra-neurite signal, A_{en}. Therefore we couple the two tissue compartments by the orientation distribution function $f(\boldsymbol{n})$. A_{en} is modelled with anisotropic (Gaussian) diffusion:

$$log A_{en} = -b \boldsymbol{q}^T \left(\int_{\mathbb{S}^2} f(\boldsymbol{n}) D(\boldsymbol{n}) \, \mathrm{d}\boldsymbol{n} \right) \boldsymbol{q}, \tag{2}$$

where $D(\boldsymbol{n})$ is a cylindrically symmetric tensor. In the present work we use the Bingham distribution as $f(\boldsymbol{n})$, as described in the next section. Hereafter the proposed model will be referred to as Bingham-NODDI, and the original as Watson-NODDI.

2.1 Bingham Distribution

The Bingham distribution [10] is a statistical parametric distribution, which is the spherical analogue of the 2-D Gaussian distribution. It quantifies the probability density of orientations along the axes in a 3-D spherical coordinate system. The distribution is characterised by three orthogonal orientations: $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$; and their respective concentrations, $\kappa_1 \geq \kappa_2 \geq \kappa_3$; as shown in Fig.1.

To simplify the representation, we redefine the concentration parameters as: $\kappa \geq \beta \geq 0$, where $\kappa = (\kappa_1 - \kappa_3)$ and $\beta = (\kappa_2 - \kappa_3)$, similarly to [11]. The Bingham distribution is then defined as:

$$f(\boldsymbol{n}) = F(\kappa, \beta)^{-1} e^{\kappa(\boldsymbol{\mu}_1 . \boldsymbol{n})^2 + \beta(\boldsymbol{\mu}_2 . \boldsymbol{n})^2} \tag{3}$$

where F is a confluent hypergeometric function.

2.2 Orientation Tensor

We summarise the orientation distribution of the neurites in each voxel in terms of an orientation tensor (OT), similar to the 3-D rendering of the diffusion tensors (DT). OT is defined as the scatter matrix (second moment) of an orientation distribution function, such as the Bingham distribution:

$$T_{i,j} = \int_{\mathbb{S}^2} \boldsymbol{n}_i f(\boldsymbol{n}) \boldsymbol{n}_j \, \mathrm{d}\boldsymbol{n} \tag{4}$$

The primary and secondary eigenvalues of the OT, $1 \geq \tau_1 \geq 1/3$, $\tau_1 \geq \tau_2 \geq 0$, are functions of the concentration parameters κ and β, and reflect the relative concentration of neurites along the dominant and primary dispersion orientations, respectively. The corresponding eigenvectors are precisely $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

Dispersion Anisotropy Index (DAI): To quantify the dispersion anisotropy of neurites, we use the planarity measure [15] of the OT: $DAI = \frac{(\tau_2 - \tau_3)}{\tau_1}$. DAI is zero for isotropic dispersion about $\boldsymbol{\mu}_1$ (Watson) and one for maximum anisotropic dispersion, i.e. when $\tau_1 = \tau_2 = 0.5$. DAI is a measure specifically related to the orientation distribution, unlike the planarity measure of a DT, which is influenced by DAI as well as other features like neurite density.

To quantify the dispersion of neurites, we use τ_1, and τ_2 as they have a finite range, unlike κ and β, which range between 0 and ∞. The concentration parameter τ_1 is inversely proportional to the dispersion parameter ODI in the original NODDI model, which is an arbitrary transformation to map κ to a finite range. To determine τ_1 and τ_2, we compute the OT from the estimated values of κ, β, using the equations 3 and 4 (The tertiary eigenvalue is by definition: $1 - \tau_1 - \tau_2$, and can be determined once τ_1 and τ_2 are computed). Since eigenvectors are mutually orthogonal, the dominant orientation, $\boldsymbol{\mu}_1$ and the primary dispersion orientation $\boldsymbol{\mu}_2$ can be quantified as a 3-D rotation of the coordinate system in which the Bingham distribution is defined. Thus we only need to estimate two extra parameters, β and the angle defining the rotation of the plane orthogonal to $\boldsymbol{\mu}_1$, compared to those estimated in Watson-NODDI.

3 Experimental Design and Results

We acquire *in vivo* dMRI data for one healthy male, on a 3T Philips scanner ($G_{max} = 60mT/m$), using the 4-shell protocol, as in [7]. We also synthesise data for a range of tissue parameters, to compare estimates against known ground truth, similarly to [7] but using the Bingham distribution.

The NODDI Matlab toolbox[1] is modified to incorporate the Bingham distribution, which is then fit to the *in vivo* and synthetic data. To assess the influence of using a simplified model to represent the neurite orientation distribution, we also fit Watson-NODDI to the data. We use Bayesian Information Criteria (BIC), a standard model selection tool, to determine which model explains the data better, while accounting for the complexity of the models. We use the optimised NODDI protocol, which is a 2-shell subset of the acquired protocol, to assess the performance of the two models. The complete 4-shell data is used as a pseudo ground-truth to calculate the errors in estimation of the parameters *in vivo*. We quantify the accuracy and precision of the parameters separately for WM and GM. For *in vivo* data, the segmentation of WM, GM and the free water compartment signal is done as described in [7]; ground-truth values of ν_{in} are used to do the same for synthetic data.

3.1 Results

Fig. 2 demonstrates the feasibility of *in vivo* estimation of neurite dispersion anisotropy. We show maps of the novel parameters, τ_1 and DAI (highlighted in yellow) obtained by fitting Bingham-NODDI, for a range of axial slices of the brain. The slices show the cross section of the corona radiata (regions in blue in the RGB map, on either sides of the corpus callosum), a region known to exhibit fanning, as it extends from the internal capsule to the various cortical areas. This fanning is captured by high values of DAI and have a good contrast to the corpus callosum, where neurites are coherently oriented. Such a contrast can not be captured by Watson-NODDI. The last two columns of Fig.2 show the maps of BIC, which clearly show that Bingham-NODDI explains the data better than Watson-NODDI without overfitting. This is consistent with the comparison of models with dispersion in [16]. Watson-NODDI performs worst specifically in areas of high dispersion anisotropy.

Table 1 shows a quantitative analysis of errors in estimation of the parameters from Bingham-NODDI and Watson-NODDI, using the 2-shell NODDI protocol. The estimation errors are very small for Bingham-NODDI estimates showing that we can accurately model dispersion anisotropy using the 2-shell clinical protocol. We also observe that the error and variability associated with $\boldsymbol{\mu_2}$ estimation is higher than those for $\boldsymbol{\mu_1}$, implying that $\boldsymbol{\mu_2}$ is harder to estimate. A key finding here is that Watson-NODDI can accurately capture neurite morphology in regions of simple neurite configurations, as the errors associated with Watson-NODDI estimates are comparable to those of Bingham-NODDI. These findings are all backed by the synthetic data analysis (results not shown).

[1] `http://nitrc.org/projects/noddi_toolbox`

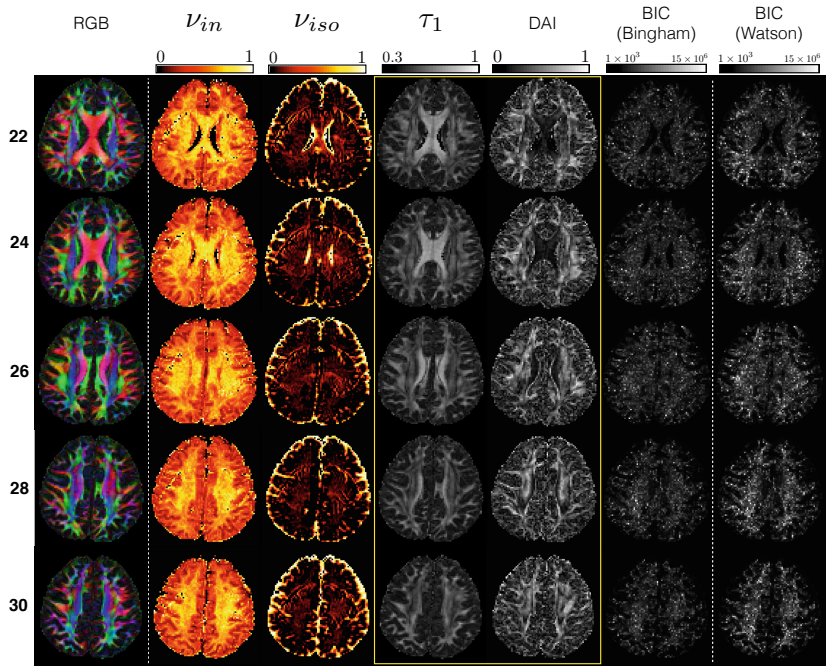


Fig. 2. Maps of the novel parameters τ_1 and DAI (highlighted in yellow), obtained from fitting Bingham-NODDI to *in vivo* data, with estimates of ν_{in}, ν_{iso} (columns 2 & 3). Corresponding RGB maps of FA weighted dominant orientation are shown for comparison. The last two columns show the maps of BIC for fitting Bingham-NODDI and Watson-NODDI.

Table 1. Mean errors with the corresponding standard deviations of those errors, for estimation of neurite parameters using Bingham-NODDI and Watson-NODDI, with respect to the 4-shell protocol estimates, for *in vivo* data

	Grey Matter		**White Matter**	
	Bingham	**Watson**	**Bingham**	**Watson**
$\boldsymbol{\nu_{in}}$	-0.025 ± 0.083	-0.025 ± 0.083	-0.002 ± 0.042	-0.001 ± 0.042
$\boldsymbol{\nu_{iso}}$	0.000 ± 0.032	0.001 ± 0.032	0.005 ± 0.029	0.005 ± 0.030
$\boldsymbol{\tau_1}$	0.006 ± 0.038	0.009 ± 0.039	-0.001 ± 0.041	0.000 ± 0.043
$\boldsymbol{\tau_2}$	0.000 ± 0.028	-0.055 ± 0.045	0.003 ± 0.029	-0.074 ± 0.047
$\boldsymbol{\mu_1}$	21.063 ± 21.458	19.709 ± 20.201	5.519 ± 7.430	5.348 ± 6.589
$\boldsymbol{\mu_2}$	27.822 ± 23.521	-	12.002 ± 14.679	-
DAI	0.011 ± 0.122	-	0.010 ± 0.078	-

4 Discussion

We demonstrate that it is possible to estimate dispersion anisotropy of neurites *in vivo*, and we can obtain sensible maps of this measure using a clinically feasible protocol. We present DAI as a measure of dispersion anisotropy which is specifically related to the orientation distribution. We show that Bingham-NODDI explains the dMRI signal better than Watson-NODDI, however presence of dispersion anisotropy does not have significant affect on the estimation of dispersion along the dominant orientation. Thus the studies based on the current implementation of NODDI are valid, but Bingham-NODDI may be used to extract parameters for dispersion anisotropy, which can enhance the findings.

A limitation of the method proposed is that it does not explicitly model crossing fibres. But our primary aim is to provide simple and robust microstructure indices and attempting to resolve multiple fibre populations will introduce instability in the model, as shown in [11]. Nevertheless, the model correctly identifies crossing regions with high orientation dispersion and some with high DAI. In future we would like to investigate the possibility of incorporating crossing fibres in Bingham-NODDI. The primary dispersion orientation is found to be harder to estimate than the dominant orientation, which is not unexpected since it represents a more subtle microstructure feature. The estimation of this feature can be improved by increasing the angular resolution of the acquisition protocol. This is not yet clinically feasible with the existing imaging sequences, but emerging technologies, such as multi-band imaging [17], will make it possible to acquire substantially more data per unit time. A reproducibility study is a natural next step for future work, as it will allow to assess if the variability in estimates is low enough to capture between-subject differences. This has implications for use of the technique in disease progression as well as tractography studies. Further work in understanding the relationship between changes in the dispersion anisotropy and normal brain development or pathology, could lead to the measure being utilised as a marker for brain disorders.

Acknowledgements. This work is supported by the EPSRC Doctoral Training Award, the MS society in the UK, and the Department of Healths NIHR Biomedical Research Centres funding scheme.

References

1. Conel, J.L.: The postnatal development of the human cerebral cortex. Harvard University Press, Cambridge (1939)
2. Jacobs, B., Schall, M., Prather, M., Kapler, E., Driscoll, L., Baca, S., Jacobs, J., Ford, K., Wainwright, M., Treml, M.: Regional dendritic and spine variation in human cerebral cortex: a quantitative Golgi study. Cerebral Cortex 11, 558–571 (2001)
3. Fiala, J.C., Spacek, J., Harris, K.M.: Review: Dendritic spine pathology: Cause or consequence of neurological disorders. Brain Research Reviews 39, 29–54 (2002)
4. Basser, P.J., Mattiello, J., LeBihan, D.L.: MR diffusion tensor spectroscopy and imaging. Biophysical Journal 66, 259–267 (1994)

5. Jespersen, S.N., Kroenke, C.D., Østergaard, L., Ackerman, J.J.H., Yablonskiy, D.A.: Modeling dendrite density from magnetic resonance diffusion measurements. NeuroImage 34, 1473–1486 (2007)
6. Jespersen, S.N., Bjarkam, C.R., Nyengaard, J.R., Chakravarty, M.M., Hansen, B., Vosegaard, T., Østergaard, L., Yablonskiy, D.A., Nielsen, N.C., Vestergaard-Poulsen, P.: Neurite density from magnetic resonance diffusion measurements at ultrahigh field: Comparison with light microscopy and electron microscopy. NeuroImage 49, 205–216 (2010)
7. Zhang, H., Schneider, T., Wheeler-Kingshott, C.A.M., Alexander, D.C.: NODDI: Practical *in vivo* neurite orientation dispersion and density imaging of the human brain. NeuroImage 61, 1000–1016 (2012)
8. Winston, G.P., Micallef, C., Symms, M.R., Alexander, D.C., Duncan, J.S., Zhang, H.: Advanced diffusion imaging sequences could aid assessing patients with focal cortical dysplasia and epilepsy. Epilepsy Research 108, 336–339 (2014)
9. Kunz, N., Zhang, H., Vasung, L., O'Brien, K.R., Assaf, Y., Lazeyras, F., Alexander, D.C., Hüppi, P.S.: Assessing white matter microstructure of the newborn with multi-shell diffusion MRI and biophysical compartment models. NeuroImage 96, 288–299 (2014)
10. Mardia, K.V., Jupp, P.E.: Directional statistics. Wiley series in probability and statistics. John Wiley & Sons, Ltd. (1990)
11. Sotiropoulos, S.N., Behrens, T.E., Jbabdi, S.: Ball and rackets: Inferring fiber fanning from diffusion-weighted MRI. NeuroImage 60, 1412–1425 (2012)
12. Jespersen, S.N., Leigland, L.A., Cornea, A., Kroenke, C.D.: Determination of axonal and dendritic orientation distributions within the developing cerebral cortex by diffusion tensor imaging. IEEE Transactions in Medical Imaging 31, 16–32 (2012)
13. Rowe, M.C., Zhang, H., Oxtoby, N., Alexander, D.C.: Beyond crossing fibers: Tractography exploiting sub-voxel fibre dispersion and neighbourhood structure. In: Gee, J.C., Joshi, S., Pohl, K.M., Wells, W.M., Zöllei, L. (eds.) IPMI 2013. LNCS, vol. 7917, pp. 402–413. Springer, Heidelberg (2013)
14. Kaden, E., Knösche, T.R., Anwander, A.: Parametric spherical deconvolution: Inferring anatomical connectivity using diffusion MR imaging. NeuroImage 37, 474–488 (2007)
15. Westin, C.-F., Maier, S.E., Mamata, H., Nabavi, A., Jolesz, F.A., Kikinis, R.: Processing and visualization for diffusion tensor MRI. Medical Image Analysis 6, 93–108 (2002)
16. Ferizi, U., Schneider, T., Tariq, M., Wheeler-Kingshott, C.A.M., Zhang, H., Alexander, D.C.: The importance of being dispersed: A ranking of diffusion MRI models for fibre dispersion using *in vivo* human brain data. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part I. LNCS, vol. 8149, pp. 74–81. Springer, Heidelberg (2013)
17. Feinberg, D.A., Moeller, S., Smith, S.M., Auerbach, E., Ramanna, S., Gunther, M., Glasser, M.F., Miller, K.L., Ugurbil, K., Yacoub, E.: Multiplexed Echo Planar Imaging for sub-second whole brain fMRI and fast diffusion imaging. PLoS One 5, e15710 (2010)

Diffusion of Fiber Orientation Distribution Functions with a Rotation-Induced Riemannian Metric*

Junning Li, Yonggang Shi, and Arthur W. Toga

Laboratory of Neuro Imaging (LONI),
Institute for Neuroimaging and Informatics (INI),
Keck School of Medicine of USC, USA
{Junning.Li,yshi,toga}@loni.usc.edu

Abstract. Advanced diffusion weighted MR imaging allows non-invasive study on the structural connectivity of human brains. Fiber orientation distributions (FODs) reconstructed from diffusion data are a popular model to represent crossing fibers. For this sophisticated image representation of connectivity, classical image operations such as smoothing must be redefined. In this paper, we propose a novel rotation-induced Riemannian metric for FODs, and introduce a weighted diffusion process for FODs regarding this Riemannian manifold. We show how this Riemannian manifold can be used for smoothing, interpolation and building image-pyramids, yielding more accurate or intuitively more reasonable results than the linear or the unit hyper-sphere manifold.

1 Introduction

The Human Connectome Project [1] has provided cutting-edge diffusion MR imaging techniques to study brain anatomical connectivity in vivo at unprecedented spatial and angular resolution. This advancement urges the development of novel mathematical methods to modelling neural fibers.

Fiber orientation distribution (FOD) images, reconstructed from diffusion weighted images, are widely used to represent the spatial and orientational distribution of neural fibers. At each voxel location p, the likelihood of fibers along a direction u is described by a real-valued and nonnegative function $F(u)$.

The mathematical property of FODs is more complicated than that of intensity values or diffusion tensors, so their fundamental operations such as smoothing and interpolation cannot be conducted in the same way. For example, linear smoothing does not yield satisfactory results, as shown in Section 5.2. In [2,3,4], $f = \sqrt{F}$ is modeled as a point on the unit hyper-sphere, a manifold whose exponential and logarithmic maps are well studied. Later in [5,6], FODs are separated as two parts: orientation and shape. FODs' comparison and interpolation are conducted by rotationally matching their shapes.

* This work is supported by grants 5U01MH093765, 7P41EB015922, 5R01MH094343 and K01EB013633 from the National Institutes of Health (NIH).

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 249–256, 2014.

In this paper, we propose a rotation-induced Riemannian metric for square-rooted FODs f, and introduce a weighted diffusion process for FODs regarding this metric. This novel weighted diffusion process can be used for smoothing, interpolation and down-sampling of FOD images. It may potentially benefit many down-stream tasks, such as tractography or registration. It can also be used to build FOD image pyramids for multi-scale image processing. Contrast to the work in [5,6], we focus on the differential structure of FODs, instead of their weighted average. In an experiment with real images, we compared the new method with the linear manifold and the hyper-sphere manifold. The new method keeps fiber integrity better in smoothing and down-sampling, and yields more accurate results in interpolation.

2 Rotation and Spherical Functions

2.1 Rotation and Angular Velocity

Any rotation in $\mathcal{R}^3$ can be described with a rotation vector $\overrightarrow{r} = [r_x, r_y, r_z]$ ($\|\overrightarrow{r}\| \leqslant \pi$) whose direction $\overrightarrow{r}/\|\overrightarrow{r}\|$ is the pivot axis according to the right-hand rule and whose length $\|\overrightarrow{r}\|$ is the rotation angle. The rotation matrix associated with $\overrightarrow{r}$ is $R = e^{[\overrightarrow{r}]^\times}$ where $[\bullet]^\times$ is the cross-product matrix of a 3D vector as defined in Eq. (1),

$$[\bullet]^\times \equiv \begin{bmatrix} 0 & -\bullet_z & \bullet_y \\ \bullet_z & 0 & -\bullet_x \\ -\bullet_y & \bullet_x & 0 \end{bmatrix}. \tag{1}$$

An angular velocity vector $\overrightarrow{\omega}$ describes the pivot axis and spinning speed of a rotating object respectively with its direction and amplitude. If an object spinning with $\overrightarrow{\omega}$ for time duration t, the accumulated rotation effect, parameterized as a rotation matrix, is $R = e^{t[\overrightarrow{\omega}]^\times}$. With R, a point $p \in \mathcal{R}^3$ rotates to Rp.

2.2 Rotation of Spherical Functions

If a spherical function $f(u)$ is rotated with R, it becomes $f(R^{-1}u)$. Given a unit angular velocity $\overrightarrow{\omega}$ ($\|\overrightarrow{\omega}\| = 1$), we define

$$\dot{f}_{\overrightarrow{\omega}}(u) \equiv \frac{d}{dt}\,|_{t=0}\, f(e^{-t[\overrightarrow{\omega}]^\times}u) \tag{2}$$

as the rotation differential of f around axis $\overrightarrow{\omega}$.

3 Manifolds of FODs

An FOD F is a spherical probability density function satisfies both $\int_{u\in S^2} F(u)du = 1$ and $F(u) \geqslant 0$, where S^2 is a unit sphere in $\mathcal{R}^3$. Cheng,

etc. in [2] modelled the manifold $M \equiv \{f(u) = \sqrt{F(u)}\}$ as a unit hyper-sphere, assuming that f can be linearly represented by a finite number of orthonormal base functions. In this section, we propose a Riemannian metric for M, differently weighting the effect of rotation in $\mathcal{R}^3$. For simplicity, symbols such as "f", "v" or "l", depending on the context, denote either spherical functions, or their linear coordinates as represented by orthonormal base functions.

3.1 Hyper-Sphere

For two spherical functions f_1 and f_2, we define $\langle f_1, f_2 \rangle \equiv \int_{u \in S^2} f_1(u) f_2(u) du$ as their dot product. As any $f \in M$ satisfies $\langle f, f \rangle = 1$, M is a unit hyper-sphere, a well-studied Riemannian manifold. Given a point $f \in M$, another point $f^\dagger \in M$ near f can be logarithmically mapped to $T_f M$, the tangent space at f, as

$$\log_f(f^\dagger) = v_f = \frac{f^\dagger - f\cos\varphi}{\|f^\dagger - f\cos\varphi\|}\varphi, \text{ where } \varphi = \arccos\left(\langle f, f^\dagger \rangle\right) \tag{3}$$

The Riemannian metric at f associated with the logarithmic map is $g_f(v_1, v_2) = \langle v_1, v_2 \rangle$ where $v_1, v_2 \in T_f M$ are two tangent vectors at f.

3.2 Rotation-Induced Riemannian Metric

Riemannian metric g_f is isotropic, treating all the directions in $T_f M$ equally. To incorporate rotation into the metric, tangent vectors generated by rotating f in $\mathcal{R}^3$ should be treated differently. Let $\dot{f}_x$, $\dot{f}_y$ and $\dot{f}_z$ denote the rotation differential of f respectively around the x, y, and z axes, as defined in Eq. (2). As $\dot{f}_x$, $\dot{f}_y$ and $\dot{f}_z$ are in $T_f M$, we use them to induce a Riemannian metric as follows. First, for a tangent vector v, we project it to the subspace spanned by $\dot{f}_x$, $\dot{f}_y$ and $\dot{f}_z$. Let $c_x\dot{f}_x + c_y\dot{f}_y + c_z\dot{f}_z$ be the projection, $c \equiv [c_x, c_y, c_z]$ the projection coefficients, and $\xi \equiv v - c_x\dot{f}_x + c_y\dot{f}_y + c_z\dot{f}_z$ the residual. Second, given two tangent vectors v_1 and v_2, the induced Riemannian metric is defined as

$$g_f^*(v_1, v_2) = \lambda \langle c_1, c_2 \rangle + \langle \xi_1, \xi_2 \rangle$$

where $\lambda > 0$ is a parameter weighting the contribution of rotation. Such a metric definition is equivalent to $g_f^*(v_1, v_2) = \langle v_1^*, v_2^* \rangle$ where $v_\bullet^* = G_f v_\bullet$ and G_f is a full-rank transformation matrix induced by $\dot{f}_x$, $\dot{f}_y$, $\dot{f}_z$ and λ. The g_f in Section 3.1 can be regarded as using $G_f = I$.

As $\dot{f}_x$, $\dot{f}_y$ and $\dot{f}_z$ are not necessarily linearly independent, the projection coefficients c may not be unique. To resolve the ambiguity, we choose the coefficients with the least norm. Let $b_1, \cdots, b_m$ be a set of unitary bases of the subspace spanned by $\dot{f}_x$, $\dot{f}_y$ and $\dot{f}_z$. The least norm projection coefficients are

$$c = A^T(AA^T)^{-1} \begin{bmatrix} \langle b_1, v \rangle \\ \vdots \\ \langle b_m, v \rangle \end{bmatrix} \text{ where } A = \begin{bmatrix} \left\langle b_1, \dot{f}_x \right\rangle & \left\langle b_1, \dot{f}_y \right\rangle & \left\langle b_1, \dot{f}_z \right\rangle \\ \vdots & \vdots & \vdots \\ \left\langle b_m, \dot{f}_x \right\rangle & \left\langle b_m, \dot{f}_y \right\rangle & \left\langle b_m, \dot{f}_z \right\rangle \end{bmatrix}.$$

The Riemannian metric g_f^* introduces a new logarithmic map near f. The length of a smooth curve $\gamma : [0,1] \to M$ is $L^*(\gamma) = \int_0^1 \sqrt{g_\gamma^*[\gamma'(t), \gamma'(t)]}dt$. The geodesic from f to a nearby point $f^\dagger$ is a curve γ which satisfies $\gamma(0) = f$ and $\gamma(1) = f^\dagger$ and minimizes $L^*(\gamma)$. With appropriate re-parameterization of t, $g_\gamma^*[\gamma'(t), \gamma'(t)]$ can be time-invariant. Accordingly, the logarithmic map of $f^\dagger$ to $T_f M$ is $\log_f^*(f^\dagger) = G_f \gamma'(0)$.

4 Weighted Diffusion of FODs

An FOD image is a mapping from $\mathcal{R}^3$ to an FOD manifold M. Given a point p in $\mathcal{R}^3$, let f_p denote its square-rooted FOD. To calculate the diffusion force at p, we construct a local helper function $h_p : h_p(q) \equiv LOG_{f_p}(f_q)$ where q is a point near p and $h_p(q)$ is the logarithmic map of f_q to $T_{f_p} M$. It should be noted that the definition of LOG_{f_p} is subject to the definition of Riemannian metrics. Sections 3.1 and 3.2 describes different metrics and logarithmic maps.

A weighted diffusion process on an FOD image is

$$\frac{df_p}{dt} = \frac{1}{w_p} l_p \equiv \frac{1}{w_p} G_{f_p}^{-1} \left(\frac{\partial^2 h_p}{\partial x^2} + \frac{\partial^2 h_p}{\partial y^2} + \frac{\partial^2 h_p}{\partial z^2} \right) \tag{4}$$

where l_p is the diffusion force at point p, and w_p is a positive weight associated with point p in the FOD image. $G_{f_p}^{-1}$ maps the diffusion force to tangent space of the hyper-sphere manifold. FODs at points with higher weights changes slower than those with lower weights. We assume that such a diffusion process eventually makes every point in the image take the same FOD value.

The weighed diffusion can be applied to FOD image processing as follows:

- Smoothing: diffusion in $\mathcal{R}^3$.
- Interpolation: weighted diffusion within a 2x2x2 voxel cube.
- Down-Sampling: diffusion followed by interpolation.

5 Evaluation

We compare three different manifold models of FODs: the linear model, the unit hyper-sphere model, and the rotation-induced manifold. For simplicity, we call them the linear, the hSphere and the rSphere methods. With the three models, image processing tasks smoothing, interpolation and down-sampling are conducted on real FOD images. Their performances are evaluated with two indices: (1) the sharpness of their output FODs and (2) the difference from baseline images depending on the processing tasks. The sharpness of FODs is measured with the generalized Fractional Anisotropy $\text{gFA}(f) = 2\text{dist}(f_0, f)/\pi$ where f_0 is the uniform spherical function in M, and $\text{dist}(\bullet, \bullet) \equiv \arccos(\langle \bullet, \bullet \rangle)$; the difference from a baseline is measured with $\delta(f) = 2\text{dist}(f_b, f)/\pi$ where f_b is the baseline FOD. (The factor $2/\pi$ scales the indices into range $[0,1]$.)

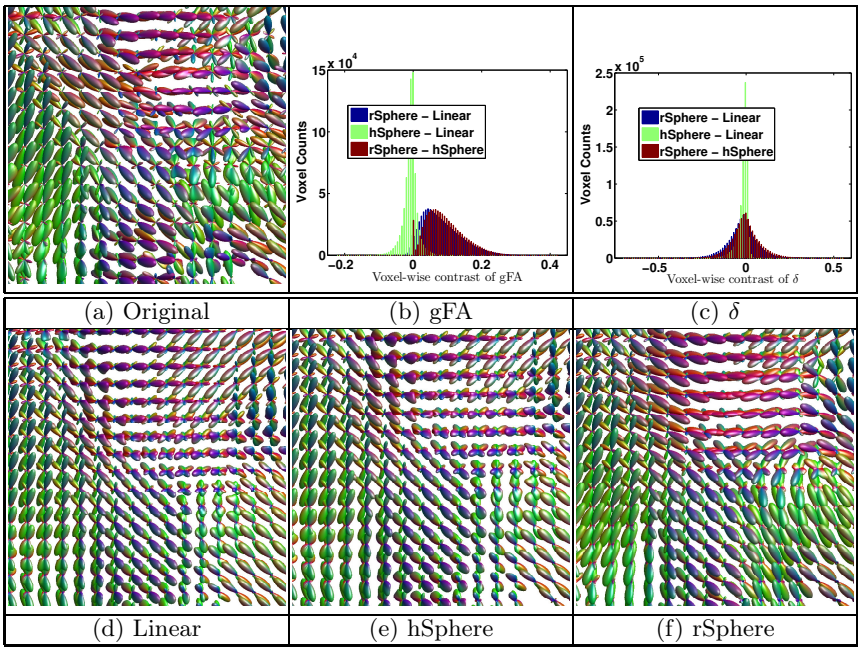


Fig. 1. Example of smoothing. The rSphere method produces sharper and orientationally smoother FODs. The linear and the hSphere methods produce "bloated" FODs.

Pixel-wise comparison within each subject is conducted as follows: a performance index image of one method is subtracted from that of another method, and then the histogram of the subtracted image is plotted. The subject-wise comparison is conducted as follows: for every subject, a voxel-averaged performance index is calculated, and the averaged indices of subjects are compared with the t-test between the three methods.

5.1 Data

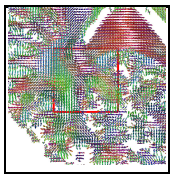
Ten subjects are randomly selected from the Human Connectome Project. The diffusion data were acquired with a multi-shell sampling scheme, and the FOD images, represented with the 8-th order spherical harmonics [7], were reconstructed with the method in [8].

5.2 Smoothing

Images are diffused with time duration $t = 1$. gFA is calculated on the diffused images, and the original images are the baseline images of δ. The diffused images of a subject is shown in Figure 1, taken from the highlighted region in the figure besides Table 1. The rSphere method yields more reasonable results, as the fibers are sharper and orientationally smoother. The linear and hSphere methods yield

Table 1. Subject-wise comparison of smoothing

	gFA		δ	
	mean	p-value	mean	p-value
rS. - hS.	0.092 ± 0.002	8.14e-16	-0.0059 ± 0.0011	6.54e-7
rS. - L.	0.084 ± 0.002	2.02e-15	-0.0221 ± 0.0013	2.22e-11
hS. - L.	-0.007 ± 0.003	1.51e-12	-0.0162 ± 0.0004	3.33e-15



"bloated" FODs. Figures 1(b) and 1(c) show the histogram of the pixel-wise contrast of the gFA and δ indices among the three methods. The rSphere method produces higher gFA than the other two methods. Table 1 shows the t-test of subject-wise comparison. The rSphere method makes slightly less changes (low δ) from the original images, yet keeps the diffused FODs much sharper (high gFA). Tiny p-values evidence the statistical significance of the results.

5.3 Interpolation

Images are sub-sampled into two parts: one of odd-index voxels and the other of even-index. FODs at even-index voxels are then interpolated from the odd-index image, and then compared with the original even-index image. gFA is calculated on the interpolated images, and the original even-index images are the baseline images of δ. The interpolation results of a subject is shown in Figure 2, taken from the highlighted region in the figure besides Table 2. The rSphere method yields more accurate results, with FODs more similar to the original image. Figures 2(b) and 2(c) show the histogram of the pixel-wise contrast of the gFA and δ indices among the three methods. The rSphere method yields higher gFA and lower δ than the other two methods. Table 2 shows the t-test of subject-wise

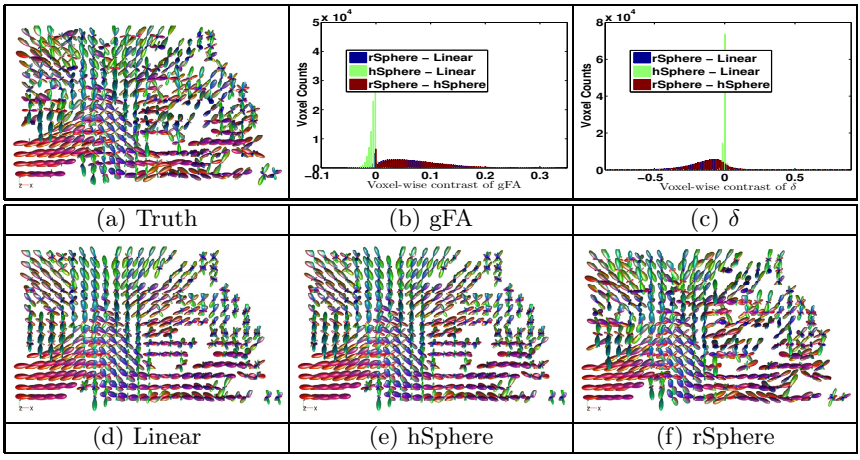
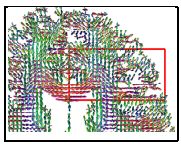


Fig. 2. Example of interpolation. The rSphere results are more similar to the truth.

Table 2. Subject-wise comparison of interpolation

	gFA		δ	
	mean	p-value	mean	p-value
rS. - hS.	$+0.071 \pm 0.0015$	3.35e-15	-0.131 ± 0.0034	1.72e-14
rS. - L.	$+0.064 \pm 0.0015$	6.87e-15	-0.134 ± 0.0035	1.95e-14
hS. - L.	-0.0061 ± 0.0002	2.60e-13	-0.0033 ± 0.0005	2.45e-07



comparison. The rSphere outperforms the other two methods with higher gFA and lower δ. Tiny p-values evidence the statistical significance of the results.

5.4 Down-Sampling

Down-sampling is conducted as follows. Images are diffused with time duration $t = 1$ and then is interpolated for the down-sampled voxel locations. Two-level down-sampling is conducted to build image pyramids. gFA is calculated on the down-sampled images. The down-sampled results of a subject is shown in Figure 3. Sub-figures (a) and (b) show respectively the overview and highlighted region of the original image. Sub-figures (c) and (e) show the down-sampled images of the hSphere method, sub-figures (d) and (f) show those of the rSphere method. (Due to limited page room, the following contents are not presented here: the results of linear down-sampling which are very similar to the hSphere method, the histogram of pixel-wise contrast of the gFA index which is very similar to the smoothing and interpolation tasks.) The rSphere produces sharper FODs. Table 3 shows the t-test of subject-wise comparison of gFA. The rSphere yields higher

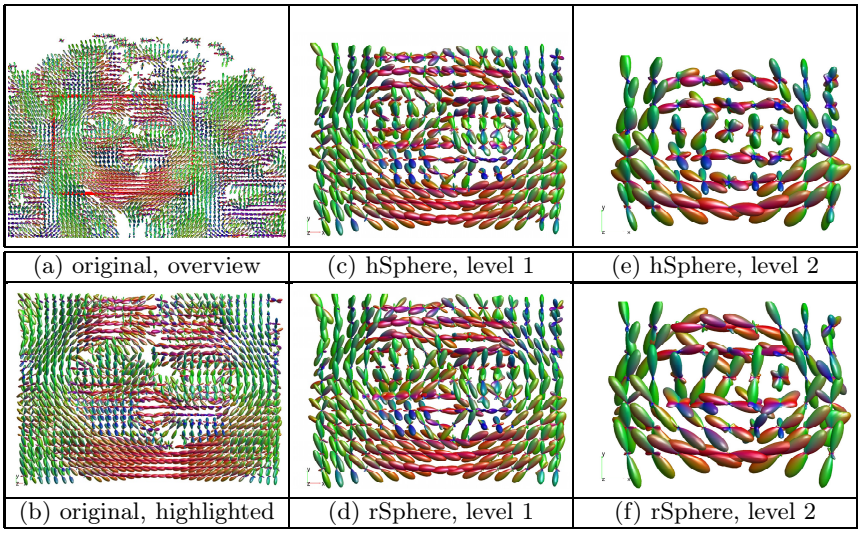


Fig. 3. Example of down-sampling. The rSphere method produces sharper FODs.

Table 3. Subject-wise comparison of down-sampling, regarding the gFA index

	Level-1		Level-2	
	mean	p-value	mean	p-value
rSphere - hSphere	$+0.035 \pm 0.0008$	3.43e-15	0.082 ± 0.0012	8.03e-17
rSphere - Linear	$+0.033 \pm 0.0008$	2.85e-15	0.075 ± 0.0011	7.82e-17
hSphere - Linear	-0.0015 ± 0.0001	1.24e-13	-0.0077 ± 0.0002	4.30e-15

gFA, indicating the down-sampled FODs are shaper than the other methods. Tiny p-values evidence the statistical significance of the results.

6 Conclusion

Weighted diffusion of square-rooted FODs f with the rotation-induced Riemannian metric leads to significantly better smoothing, interpolation and down-sampling of FOD images, in comparison with the linear manifold, and the hypersphere manifold.

References

1. Human Connectome Project, `http://www.humanconnectomeproject.org/`
2. Cheng, J., Ghosh, A., Jiang, T., Deriche, R.: A Riemannian framework for orientation distribution function computing. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009, Part I. LNCS, vol. 5761, pp. 911–918. Springer, Heidelberg (2009)
3. Du, J., Goh, A., Qiu, A.: Diffeomorphic metric mapping of high angular resolution diffusion imaging based on Riemannian structure of orientation distribution functions. IEEE Transactions on Medical Imaging 31(5), 1021–1033 (2012)
4. Srivastava, A., Jermyn, I., Joshi, S.: Riemannian analysis of probability density functions with applications in vision. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8. IEEE (2007)
5. Ncube, S., Srivastava, A.: A novel Riemannian metric for analyzing HARDI data. In: SPIE Medical Imaging, p. 79620Q. International Society for Optics and Photonics (2011)
6. Cetingul, H.E., Afsari, B., Wright, M.J., Thompson, P.M., Vidal, R.: Group action induced averaging for HARDI processing. In: 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1389–1392. IEEE (2012)
7. Hill, D.A.: Electromagnetic Fields in Cavities: Deterministic and Statistical Theories. Wiley-IEEE Press (2009)
8. Tran, G., Shi, Y.: Adaptively constrained convex optimization for accurate fiber orientation estimation with high order spherical harmonics. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part III. LNCS, vol. 8151, pp. 485–492. Springer, Heidelberg (2013)

Machine Learning Based Compartment Models with Permeability for White Matter Microstructure Imaging

Gemma L. Nedjati-Gilani[1], Torben Schneider[2] Matt G. Hall[1], Claudia A.M. Wheeler-Kingshott[2], and Daniel C. Alexander[1]

[1] Centre for Medical Image Computing and Dept. of Computer Science, University College London, Gower St, London, UK
g.nedjati-gilani@ucl.ac.uk
[2] Dept. of Neuroinflammation, Institute of Neurology, University College London, Queen Square, London, UK

Abstract. The residence time τ_i of water inside axons is an important biomarker for white matter pathologies of the human central nervous system, as myelin damage is hypothesised to increase axonal permeability, and thus reduce τ_i. Diffusion-weighted (DW) MRI is potentially able to measure τ_i as it is sensitive to the average displacement of water molecules in tissue. However, previous work addressing this has been hampered by a lack of both sensitive data and accurate mathematical models. We address the latter problem by constructing a computational model using Monte Carlo simulations and machine learning in order to learn a mapping between features derived from DW MR signals and ground truth microstructure parameters. We test our method using simulated and in vivo human brain data. Simulation results show that our approach provides a marked improvement over the most widely used mathematical model. The trained model also predicts sensible microstructure parameters from in vivo human brain data, matching values of τ_i found in the literature.

1 Introduction

Numerous white matter (WM) pathologies of the human central nervous system (CNS), such as multiple sclerosis, spinal cord injury and leukodystrophies, are characterised by damage to the insulating myelin sheaths around the axonal fibres. As the breakdown of myelin is hypothesised to lead to an increase in axonal permeability, there is widespread interest in developing imaging biomarkers based on permeability or intra-axonal water residence time τ_i in order to improve diagnosis of and treatment for these conditions. Diffusion-weighted (DW) MRI is potentially amenable to estimating τ_i as it is sensitive to the dispersion of water molecules within tissue. However, due to a lack of both sensitive data and sufficiently accurate mathematical models, progress has been limited. Whereas improvements in modern hardware and the development of specialised imaging sequences are beginning to address the former issue, there is still a need

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 257–264, 2014.

to develop models that accurately characterise water exchange within the brain. Many commonly used compartment models for microstructure imaging, such as AxCaliber [4] and ActiveAx [3], ignore the effects of permeability completely and overestimate axon diameter. Including the effects of permeability in these techniques not only gives us another clinically useful parameter τ_i, but may also improve the estimation of other microstructure indices.

Mathematical models that explicitly incorporate τ_i include the Kärger model (KM) and apparent exchange rate (AXR) imaging. The KM is commonly used [8,12] as it is compatible with data acquired using widely available pulsed gradient spin echo (PGSE) and stimulated echo (STE) imaging sequences. It accounts for intercompartmental water exchange by coupling the DW MR signals due to the separate compartments via τ_i. However, it relies on the assumption that the two pools of water are well mixed and it does not model restriction. These conditions are not valid in WM tissue where the intra- and extra-axonal compartments are spatially separate and the axonal membranes restrict the motion of water molecules. Even though these limitations have been known for over 20 years, there have been no improvements to the model due to the mathematical intractability of the problem. AXR imaging [9] has recently been introduced as an alternative to the KM; however it requires specialised double PGSE imaging sequences, and it again relies on strong assumptions about the compartmentation of water into a 'fast' and 'slow' pool. The estimated AXR parameter also conflates τ_i with intra-axonal volume fraction f, making it difficult to disentangle the origin of any measured change.

Given the inherent difficulties involved in deriving accurate analytical models of permeability, we approach this problem by constructing a computational model to learn the mapping between microstructural features of interest and the data, bypassing the need for a mathematical model altogether. We use Monte Carlo (MC) simulations to generate synthetic signals from a library of histologically relevant microstructure indices. A random forest (RF) regressor then learns the mapping between features derived from DW MR signals and ground truth microstructure parameters using the synthetic data, providing an efficient and accurate method for predicting microstructure parameters, including τ_i.

Previous related work [8] generated libraries of microstructure parameters and their corresponding DW MR signals from MC simulations, and used them to find the nearest-neighbour microstructure parameters that matched unseen signals; however nearest-neighbour matching typically has poor generalisation to unseen input data and the method has only been demonstrated on synthetic data. We extend this approach using RF regression which has better generalisation to unseen data [5], i.e. combinations of tissue parameter values not explored in the training set. We compare our approach to the KM using simulated data and demonstrate that the trained RF can be used to predict sensible estimates of microstructure indices from in vivo human brain WM.

2 Methods

Protocol Optimisation. We use an orientationally invariant (OI), DW-STE protocol, previously optimised [2] for a two-compartment KM assuming a maximum imaging time of 30 minutes. The protocol was optimised for the following biophysically plausible tissue parameters: f=0.7, parallel diffusivity $d_{\parallel}$=2×10^{-9} m^2s^{-1}, perpendicular diffusivity $d_{\perp}$=0.7×10^{-9} m^2s^{-1}, axon radius R=1 μm, $\tau_i\in\{0.05,\ 0.1,\ 0.2,\ 0.4,\ 1\}$ s, $T1$=832 ms. The resulting protocol contains 89 measurements divided into 4 distinct shells, with Δ ranging from 95 to 398 ms. The final protocol, accounting for the effects of the additional STE imaging gradients, is shown in table 1.

Monte Carlo Simulations. We use MC simulations [6], in combination with the OI protocol in table 1, to generate DW MR signals from 12,500 WM tissue substrates. We model WM as a collection of 100,000 non-abutting, parallel cylinders with radii drawn from a gamma distribution (with mean μ_R, standard deviation σ_R). Each substrate is described by a unique combination of μ_R, σ_R, f, τ_i and $d_{\parallel}$ which are randomly selected in the ranges: $\mu_R \in [0.2, 5]$ μm, $\sigma_R \in[\min(0.1,\frac{\mu_R}{5}),\ \frac{\mu_R}{2}]$ μm (to ensure that the distributions have a non-zero mode, matching the distributions observed in histology [1]), $f \in [0.4, 0.7]$, $\tau_i \in [20, 950]$ ms, $d_{\parallel} \in [0.8, 2.2]\times10^{-9}$ m^2s^{-1}. All simulations are performed using 100,000 spins and 2,000 time steps. We generate two sets of signals: noise-free and noisy. As spins undergo $T1$ relaxation during the mixing time TM between the two diffusion gradients, measurements made using longer Δ (and so longer TM) experience more relaxation leading to lower signal intensities and signal to noise ratios (SNR). For the noisy data set we scale the signals by $\exp\left(-\frac{TM}{T1}\right)$ where $T1$=832 ms for WM at 3T. We then add Rician noise, choosing the standard deviation of the noise σ so that the SNR of the $b = 0$ images with Δ=95 ms is 20.

Data Acquisition and Pre-processing. We acquire DW images, using the protocol in table 1, from one healthy subject (male, age 31) on a 3T Phillips Achieva scanner with the following imaging parameters: TE=55 ms, TR=12000 ms, FOV=256mm×256mm, matrix size=128×128, slice thickness=4 mm, no. of slices=30. The total imaging time is approximately 30 minutes. As the model

Table 1. OI STE protocol parameters, optimised for Kärger model parameter estimation

Shell	# $b = 0$ (s mm^{-2})	# gradient directions	b (s mm^{-2})	$\lvert G\rvert$ (mT m^{-1})	Δ (ms)	δ (ms)	TM (ms)
1	1	32	1592	70	95	7	69
2	1	6	1746	62	398	4	370
3	1	32	3538	71	394	5	370
4	1	15	3950	84	162	7	135

that we learn here is specific to axons that resemble parallel cylinders, it is not applicable in regions containing CSF, grey matter or highly dispersed or crossing WM fibres. We mask out these voxels by computing maps of linearity $C_L = \frac{\lambda_1 - \lambda_2}{\lambda_1}$ and planarity $C_P = \frac{\lambda_2 - \lambda_3}{\lambda_1}$ [13] from diffusion tensor (DT) fits, and select only those voxels with $C_L > 0.5$ and $C_P < 0.3$. The SNR of the selected WM region in the $b = 0$ image with $\Delta = 95$ ms is approximately 19.

Random Forest Regression. RF regression works by averaging the predictions from an ensemble of randomly trained decision trees [5]. We use it here to learn a mapping between rotationally invariant features derived from DT (all shells) and 4th order spherical harmonic (SH) fits (shells 1,3,4) to the simulated DW MR signals and the ground truth microstructure parameters. From the DT fits we calculate the eigenvalues $\lambda_1, \lambda_2, \lambda_3$, mean diffusivity and fractional anisotropy. From the SH fits we calculate the mean, peak, dispersion (i.e. the eigenvalues of the hessian matrix at the peak), anisotropy, skewness and kurtosis of the apparent diffusion coefficient profile, as well as simple combinations of the SH coefficients given by $I_k = \sum_{i=-k}^{k} |a_{k,i}|^2$ for $k = 0, 2, 4$ where $a_{k,i}$ is the coefficient of SH order k and index i. This gives a vector with 50 features for each measurement. The RF regressor, containing 100 trees of maximum depth 20, is trained [10] on 10,000 of the 12,500 feature vectors from noisy and noise-free data separately. The remaining 2,500 previously unseen feature vectors are used for testing. When predicting the microstructure indices from the noise-free test sets, we use the RF trained on noise-free data; when predicting from the noisy test and in vivo feature vectors (which have similar noise characteristics), we use the RF trained on noisy data. The microstructure parameters we estimate during the RF regression are f, $d_\parallel$, τ and α, a single axon radius index [3] which reflects both the mean and spread of the radius distribution.

Kärger Model Fitting. We fit a two-compartment KM to the 2,500 noise-free and noisy test data sets, and to the masked WM voxels from the in vivo human data set. The intra-axonal compartment, with volume fraction f, is modelled using randomly packed, parallel cylinders with radius R and an intra-axonal water residence time τ_i. The extracellular space is modelled as a cylindrically symmetric DT with diffusivities, $d_\parallel$ and $d_\perp$. $d_\parallel$ is assumed to be the same in both compartments. The model is simplified using the tortuosity model as in [3]. Prior to model fitting, each measurement is normalised by the $b = 0$ measurement with the same TM to eliminate $T1$ effects. The model is fit using Markov Chain Monte Carlo (MCMC) with an offset Gaussian noise model (assuming different noise standard deviations σ for each shell of data, which we estimate a priori) to sample from the posterior distribution over the model parameters. The burn-in phase for the MCMC consists of 10,000 steps, after which we collect 1,000 samples at an interval of 100 steps.

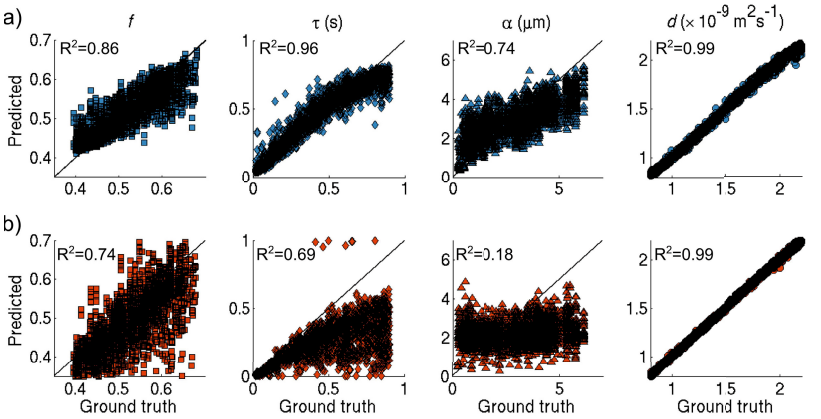


Fig. 1. Scatter plots of ground truth values of f, τ, α, $d_{\|}$ against a) predictions from the RF and b) the KM for noise-free data with correlation coefficients shown

3 Results

Figure 1 shows scatter plots of f, α, τ_i and $d_{\|}$ against predictions from a) the RF regressor and b) the KM in the case of noise-free simulated data. The correlations between ground truth and RF predictions are strong for all parameters. Even so, we do not get a perfect recovery, in part because of the statistical nature of the model, but also because parameters such as τ_i and α have a very weak influence on the DW MR signals and thus the features we derive from them. However, this provides an indication of the best predictions we can make given the measurements we have. For the KM, we observe good correlations for f and $d_{\|}$, but virtually no sensitivity to α. The protocol was optimised for sensitivity to τ_i rather than α, so this is in line with expectations. The KM can estimate τ_i for residence times less than 200 ms as, due to the long Δ used in the protocol, the intra- and extra-axonal compartments appear well mixed at this timescale for smaller τ_i; beyond this, the assumption breaks down and the KM severely underestimates large τ_i.

Figure 2 shows similar scatter plots, but for noisy data. Predictions from the RF are noisier, resulting in lower correlations with the ground truth values. Estimates of $d_{\|}$ are still strong, but large f in particular are consistently underestimated. There is still a positive correlation between ground truth and predicted values of τ_i until ≈400-500 ms, after which the estimates level off. This is because the small differences in the features due to τ_i are now overshadowed by the differences in the features due to noise. However, the RF is still able to distinguish short and long τ_i, unlike the KM, which is the key requirement for a useful imaging biomarker. The correlation for α is again much lower, although there still a slight positive trend indicating that the RF may be able to distinguish between large and small α. The KM generally provides good estimates of f and $d_{\|}$ but has almost no sensitivity to α and τ_i. Given that the KM fails to

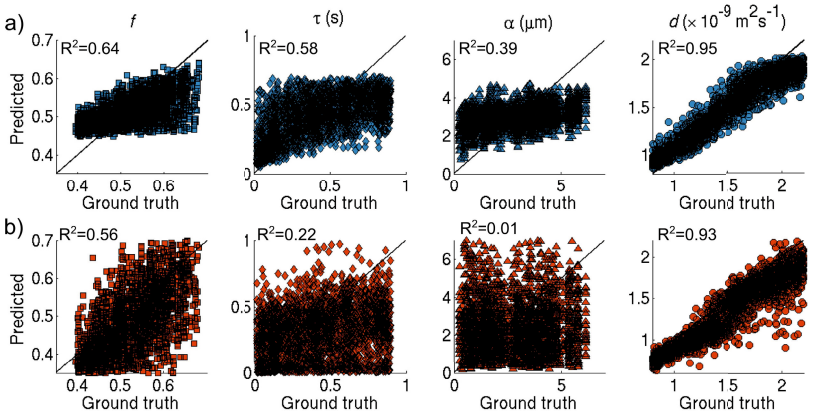


Fig. 2. Scatter plots of ground truth values of f, τ, α, $d_{\|}$ against a) predictions from the RF and b) the KM for noisy data with SNR=20, with correlation coefficients shown

estimate these parameters from data generated using a simple geometric model and realistic SNR, it is unlikely to be sensitive to these parameters in real brain tissue, which is much more complex.

We note that there is bias in this comparison, as the test data for the RF is generated in the same manner as the training data, which is not the case for the KM. However, it clearly illustrates the limitations of the KM's assumptions in situations where they are known to be violated, as in WM.

Figure 3 shows in vivo estimates of human WM microstructure parameters using a) the RF and b) the KM across the same sagittal and axial slices. The values of $d_{\|}$ estimated by both models are consistent across the WM. Predictions of f using the RF are lower than expected for WM, ranging from 0.45-0.63, reflecting the results from the noisy simulated data. Across the mid-sagittal corpus callosum (CC), we see slightly higher f in the genu and splenium compared to the midbody [1], but the trend is not as clear as that predicted by the KM. Estimates of α and τ_i from the KM are very noisy, as expected, and show no obvious patterns across the WM. In contrast, the values of α and τ_i predicted by the RF are much less noisy and we can identify trends across WM tracts. For example, across the mid-sagittal CC, we see the characteristic low-high-low trend in α [1]. Estimates of τ_i are consistently in the range 400-600 ms across the genu and splenium of the CC, and slightly lower in the corticospinal tract (CST), where we predict τ_i=300-500 ms. It is inherently difficult to validate these values as accurate estimates of τ_i are not obtainable via histology. However, a study of intra- to extra-axonal water exchange across the whole in vivo rat brain using relaxometry and contrast agents suggests a mean τ_i of approximately 550 ms [11]. This is similar to the values predicted here, although how well these numbers correspond to human tissue is unknown.

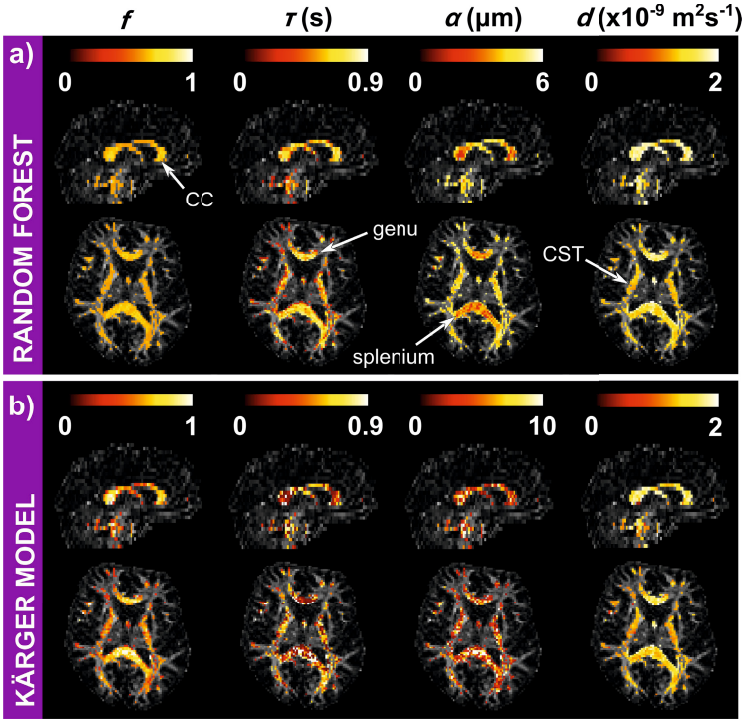


Fig. 3. Predicted values of f, τ, α, $d_{\|}$ from in vivo human brain WM across slice x=66, $z = 15$ using a) the RF regressor and b) the KM

4 Discussion

This study demonstrates that we can learn a mapping between microstructure parameters and simulated DW MR signals using RF regression, even when the data is noisy. The parameter correlations, particularly for τ_i and α, are higher for the RF than the KM, even though acquisition protocol we use was optimised for the latter approach. Furthermore, the trained RF predicts sensible microstructure parameters from in vivo human data, even for parameters such as α and τ_i which only weakly influence the DW MR signals available from human scanners. The model gives a way of obtaining plausible, if noisy, estimates of τ_i in vivo for the first time. Given the mathematical difficulty of deriving accurate analytical models of permeability, this approach is very promising; however further work and validation is needed.

Although the mapping we learn here is specifically for randomly packed, parallel, non-abutting cylinders and a STE imaging sequence, the approach can be easily extended to other tissue configurations and pulse sequences. In future we plan to incorporate models of fibre dispersion into our MC simulations, e.g. by using undulating cylinders, allowing us to extend our technique to dispersed white matter fibre regions as well as grey matter. We also plan to investigate

more specialised pulse sequences, such as the AXR sequence [9], which may improve sensitivity as well as allowing us to compare other analytic models.

The protocol used here was optimised specifically for the KM. It is therefore unlikely that this protocol is optimal for our approach, especially as some of the DW shells do not have high enough angular resolution to support feature calculation. Improving the angular resolution of the data, adding a longer diffusion time (subject to SNR constraints) and increasing $|G|$ should improve the RF predictions, particularly of τ_i and α.

Acknowledgments. This work is supported by EPSRC grant EP/I027084/1.

References

1. Aboitiz, F., Scheibel, A.B., Fisher, R.S., Zaidel, E.: Fiber composition of the human corpus callosum. Brain Res. 598, 143–153 (1992)
2. Alexander, D.C.: A general framework for experiment design in diffusion MRI and its application in measuring direct tissue-microstructure features. Mag. Res. Med. 60, 439–448 (2008)
3. Alexander, D.C., Hubbard, P.L., Hall, M.G., Moore, E.A., Ptito, M., Parker, G.J., Dyrby, T.B.: Orientationally invariant indices of axon diameter and density from diffusion MRI. Neuroimage 52, 1374–1389 (2010)
4. Assaf, Y., Blumenfeld-Katzir, T., Yovel, Y., Basser, P.J.: AxCaliber: a method for measuring axon diameter distribution from diffusion MRI. Mag. Res. Med. 59, 1347–1354 (2008)
5. Criminisi, A., Shotton, J., Konukoglu, E.: Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. Microsoft Research technical report (2011)
6. Hall, M.G., Alexander, D.C.: Convergence and parameter choice for Monte-Carlo simulations of diusion MRI. IEEE Trans. Med. Im. 28, 1354–1364 (2009)
7. Kärger, J., Pfeifer, H., Wilfried, H.: Principles and application of self-diffusion measurements by nuclear magnetic resonance. Adv. Mag. Res. 12, 1–89 (1988)
8. Nilsson, M., Alerstam, E., Wirestam, R., Ståhlberg, F., Brockstedt, S., Lätt, J.: Evaluating the accuracy and precision of a two-compartment Kärger model using Monte Carlo simulations. J. Mag. Res. 206, 59–67 (2010)
9. Nilsson, M., Lätt, J., van Westen, D., Brockstedt, S., Lasič, S., Ståhlberg, F., Topgaard, D.: Noninvasive mapping of water diffusional exchange in the human brain using filter-exchange imaging. Mag. Res. Med. 69, 1573–1581 (2013)
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., Pedregosa, F., et al.: Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011)
11. Quirk, J.D., Bretthorst, G.L., Duong, T.Q., Snyder, A.Z., Springer, C.S., Ackerman, J.J.H., Neil, J.J.: Equilibrium water exchange between the intra- and extracellular spaces of mammalian brain. Mag. Res. Med. 50, 493–499 (2003)
12. Stanisz, G.J., Wright, G.A., Henkelman, R.M., Szafer, A.: An analytical model of restricted diffusion in bovine optic nerve. Mag. Res. Med. 37, 103–111 (1997)
13. Westin, C.-F., Maier, S.E., Mamata, H., Nabavi, A., Jolesz, F.A., Kikinis, R.: Processing and visualization for diffusion tensor MRI. Med. Im. Analysis 6, 93–108 (2002)

Probabilistic Shortest Path Tractography in DTI Using Gaussian Process ODE Solvers

Michael Schober[1], Niklas Kasenburg[1,2], Aasa Feragen[1,2], Philipp Hennig[1], and Søren Hauberg[3]

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2] Department of Computer Science, University of Copenhagen, Denmark
[3] Technical University of Denmark

Abstract. Tractography in diffusion tensor imaging estimates connectivity in the brain through observations of local diffusivity. These observations are noisy and of low resolution and, as a consequence, connections cannot be found with high precision. We use probabilistic numerics to estimate connectivity between regions of interest and contribute a Gaussian Process tractography algorithm which allows for both quantification and visualization of its posterior uncertainty. We use the uncertainty both in visualization of individual tracts as well as in heat maps of tract locations. Finally, we provide a quantitative evaluation of different metrics and algorithms showing that the adjoint metric [8] combined with our algorithm produces paths which agree most often with experts.

1 Introduction

Diffusion tensor imaging (DTI) is a non-invasive imaging technology generating a global mapping of local brain diffusivity. DTI estimates, in each voxel, the local directional probability of water molecule displacement, represented as a tensor field. Assuming high diffusion along brain fibers, global connections can be estimated by integrating the tensor field, a process referred to as *tractography*.

Shortest path algorithms for tractography are useful for quantifying structural brain connectivity between regions of interest (ROIs), as needed for structural brain connectivity graphs. Due to low resolution and a low signal-to-noise ratio, one cannot find connections with high precision. Quantification and interpretation of the uncertainty of solutions is thus an important problem [18]. We use probabilistic numerics [12] both to estimate connectivity in the form of tracts connecting ROIs, and to quantify and visualize the tracts and their uncertainty.

We apply a recent algorithm [12] to infer a distribution of possible solutions to shortest path problems in tractography. This distribution is represented as a Gaussian Process (GP) over the solution to an ordinary differential equation (ODE). GPs offer novel ways to quantify and visualize uncertainty arising from the numerical computation, and allow marginalization over a space of feasible solutions. This has two strong advantages: 1) the inclusion of quantified numerical uncertainty into the final tractography result gives a more honest view of the accuracy of the algorithm; and 2) visualizing the uncertainty emphasizes the

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 265–272, 2014.

fact that the shortest paths solutions found are *not* real fibers in the brain, but mathematical abstractions of probable connectivity.

Contribution and Organization. In Sec. 1.1 we provide a brief overview of tractography. We then present the suggested algorithm for tractography based on a probabilistic numerical solution to the ODEs arising when a Riemannian metric is estimated from the diffusion tensors (Sec. 2). This gives a distribution of possible shortest paths in the form of a GP. We then show new visualization techniques emerging from the availability of the uncertainty of the GP solution (Sec. 3). Finally, the same section provides results from a quantitative analysis demonstrating that the adjoint metric [8] combined with GP ODE solvers produces shortest paths matching best to expert annotations compared to state-of-the-art in shortest path tractography.

1.1 Background and Related Work

Current tractography methods are either *fiber tracking* or *shortest path* methods.

Fiber tracking methods greedily trace the most probable path in any direction from a chosen seed until a stopping criterion is met. *Deterministic* fiber tracking algorithms [2] trace paths by following a non-stochastic rule, while *probabilistic* algorithms [15] randomly sample different directions in each step and continue forward from the best one(s).

While popular, these methods suffer from two problems. 1) Exploratory tracking methods continue tracing out a path until the algorithm no longer knows where to go. Because of this, many tracking algorithms either get lost in low-connected areas or in areas with crossing fibers. 2) Voxels near the starting region are explored more thoroughly than voxels far away. Some parts of the brain can, thus, be under-explored. This not only introduces a bias towards paths close to the starting region, but may also have the effect that the optimal path is never explored. This is particularly important when tractography is used to generate structural connectivity networks between ROIs for population studies of brain networks. Weak connections will be far less reliably estimated than strong ones, meaning that any graph analysis performed on the resulting network is biased towards finding differences in strong connections.

Shortest path methods tackle these problems by specifically computing connections between ROIs rather than connections from a seed. To derive efficient algorithms, optimal connections are often defined as *shortest paths* under a Riemannian metric, which is inversely proportional to the local diffusion tensor [13,14]. This way, local steps of short length correspond to local steps of large diffusion. While multiple shortest paths might exist between two given voxels, only one solution is obtained. Hence, many voxels are considered (cf. Fig. 3). In more detail, from the diffusion tensor D_i at voxel i, a local distance measure $\mathrm{dist}^2(a,b) = \alpha_i(a-b)^\intercal D_i^{-1}(a-b)$ is defined. Here α_i is a per-voxel scaling factor, where most commonly $\alpha_i = 1$ [13,14]. Hao et al. [11] argue that a per-voxel scaling is needed and provide a numerical scheme for computing α_i. Fuster et al. [8] show that shortest paths only correspond to free Brownian motion under the

implied Riemannian metric when $\alpha_i = \det D_i$. This is called the *adjoint metric*. We investigate this choice further in the empirical evaluation (Sec. 3).

Discrete shortest path methods construct a graph with vertices corresponding to voxel positions, edges between neighboring voxels, and edge length defined according to the local distance measure. Shortest paths can then be found either using *Dijkstra's algorithm* [14] or *fast marching methods* [16]. This is computationally efficient, but suffers from discretization errors as the shortest paths are restricted to go through the voxels.

Continuous shortest path methods smoothly interpolate the per-voxel metric tensors $M_i = \alpha_i D_i^{-1}$ to form a continuous metric space. The shortest path from a to b can then be found by solving the following system of ODEs [13]

$$\begin{aligned} c''(t) &= -\frac{1}{2} M^{-1}(c(t)) \left[\frac{\partial \operatorname{vec} M(c(t))}{\partial c(t)} \right]^{\mathsf{T}} [c'(t) \otimes c'(t)] \\ &=: f(c(t), c'(t)), \qquad \text{with } c(0) = a, \;\; c(1) = b, \end{aligned} \tag{1}$$

where $\operatorname{vec} M \in \mathbb{R}^9$ stacks the columns of M and $\otimes$ is the Kronecker product. The solution is a continuous curve $c : [0, 1] \to \mathbb{R}^3$, so discretization issues are avoided. Unlike discrete methods, the solution found by the numerical solver might only provide a locally optimal path. This concern, along with their ease of implementation, has made discrete methods the most common solution in shortest path tractography [8, 11, 14, 16].

Non-trivial ODEs such as (1) cannot be solved analytically. Traditional numerical algorithms return a point estimate in the solution space of all curves agreeing with the *finitely many* evaluation points of the ODE. This might lead to the impression that this is *the* true solution rather than an *uncertain* approximation. GP ODE solvers return a more accurate picture in this regard.

2 Methodology

We propose a probabilistic continuous shortest path tractography algorithm. Similarly to Lenglet et al. [13] we form a continuous Riemannian metric through trilinear interpolation of the metric tensors given at voxel locations. We then provide a method for solving Eq. (1) numerically which gives explicit estimates of the uncertainty of the solution.

2.1 Probabilistic ODE Solvers

The central idea in recent work on probabilistic solvers for ODEs is to assign a prior probability distribution over possible solutions to the ODE, then iteratively refine it to a posterior distribution by repeatedly evaluating the ODE at test inputs. For algebraic as well as conceptual reasons [4], GPs are a preferable choice for the prior distribution. We first review GPs and then discuss their application in ODE solvers.

Review of GP Regression. A Gaussian Process $\mathcal{GP}(c; \mu, k)$ [17] is a probability distribution over real-valued functions $c : \mathbb{R} \mapsto \mathbb{R}$ such that any finite restriction to function *values* $\{c(t_1), \ldots, c(t_N)\}$ has a Gaussian distribution. GPs

are parameterized by a mean function $\mu : \mathbb{R} \mapsto \mathbb{R}$ and a covariance function $k : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$. GPs are closed under linear transformations; a linear operation Φ induces a distribution over Φc as $\mathcal{GP}(\Phi c; \Phi\mu, \Phi k \Phi^\intercal)$, where $\Phi^\intercal$ denotes application of Φ to the left. Given observations $(T, Y) = [(t_1, y_1), \ldots, (t_N, y_N)]^\intercal$ of likelihood $\mathcal{N}(y; \Phi c, \sigma^2 \mathbf{I})$, the posterior over c is a $\mathcal{GP}(c; \tilde{\mu}, \tilde{k})$ with

$$\tilde{\mu}(t) := \mu(t) + k(t,T)\Phi^\intercal(\Phi k_{TT}\Phi^\intercal + \sigma^2\mathbf{I})^{-1}(Y - \Phi\mu(T)) \tag{2}$$

$$\tilde{k}(t,u) := k(t,u) - k(t,T)\Phi^\intercal(\Phi k_{TT}\Phi^\intercal + \sigma^2\mathbf{I})^{-1}\Phi k(T,u), \tag{3}$$

where $(k_{TT})_{ij} := k(t_i, t_j)$ is the covariance matrix of input locations [17, §2.2]. Fig. 1 illustrates the concept with a GP posterior belief $\mathcal{GP}(c; \tilde{\mu}, \tilde{k})$ (green) and its derivative $\mathcal{GP}(\partial c; \partial\tilde{\mu}, \partial\tilde{k}\partial^\intercal)$ (orange). Bold lines show posterior mean and filled areas show point-wise two times the standard deviation. The posterior was generated from 5 observations (black). Beliefs over multi-output functions $c = [c_i(t)], i = 1, \ldots, D$ can be constructed through vectorization. If the covariance structure is assumed to factorize between inputs and outputs

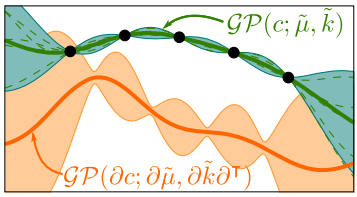


Fig. 1. GP example

$$\mathrm{cov}(c_i(t), c_j(u)) = V_{ij} \cdot k(t,u), \tag{4}$$

then the belief over c can be written as $p(c(t)) = \mathcal{GP}(c; \mu_c, V \otimes k)$. (The important special case $V = \mathbf{I}$ amounts to independent GP priors for each dimension of c). The covariance structure determines the regularity and uncertainty of the curve.

GP ODE Solvers. In [12], Hennig & Hauberg studied a framework, originally envisioned by Skilling [19], for obtaining a posterior probability distribution $p(c)$ over the solution c to an ODE like Eq. (1). The general concept was also recently analyzed by Chkrebtii et al. [4]. It works as follows.

Since differentiation is a linear operation, a GP distribution on c induces a GP belief on its derivatives, if the mean and kernel functions are sufficiently smooth. One can now repeatedly construct an estimate $\tilde{c}(t_i)$ for the true solution $c(t_i)$ to the ODE and use it to construct approximate observations $y_i = f(\tilde{c}(t_i), \partial\tilde{c}(t_i))$ of $\partial^2 c(t_i)$. As $\tilde{c}(t_i)$ is only an approximation to the true solution $c(t_i)$, the observation y_i is imprecise up to some error that, in this framework, is estimated by propagating the current uncertainty over c through the ODE:

At step i, assume a current posterior $p(c) = \mathcal{GP}(c; \tilde{\mu}^i, \tilde{k}^i)$. We construct the estimates $\tilde{c}$ and $\partial\tilde{c}$ as the current "best guess", the mean: $[\tilde{c}, \partial\tilde{c}] = [\tilde{\mu}(t_i), \partial\tilde{\mu}(t_i)]$. The estimate for the error of this guess is the current marginal variance

$$\mathrm{cov}\begin{pmatrix}\tilde{c}^i(t_i)\\ \partial\tilde{c}^i(t_i)\end{pmatrix} = \begin{pmatrix}\tilde{k}^i(t_i,t_i) & \left.\frac{\partial\tilde{k}^i(t_i,t)}{\partial t}\right|_{t=t_i}\\ \left.\frac{\partial\tilde{k}^i(t,t_i)}{\partial t}\right|_{t=t_i} & \left.\frac{\partial^2\tilde{k}^i(t,u)}{\partial t\partial u}\right|_{t,u=t_i}\end{pmatrix} =: \Sigma^i. \tag{5}$$

Assuming we have access to upper bounds or other estimates $U > {}^{\partial f}/_{\partial c}$ and $U' > {}^{\partial f}/_{\partial c'}$ on the gradients of f, we can use Σ^i to construct an estimate

for the error on y_i as $[U, U']^\intercal \Sigma^i [U, U'] =: \Lambda^i$, which gives a likelihood function $p(y_i \mid \partial^2 c(t_i)) = \mathcal{N}(y_i; f(\tilde{c}^i, \partial\tilde{c}^i), \Lambda^i)$. Using (2) and (3), the belief can be updated to obtain $\tilde{\mu}^{i+1}, \tilde{k}^{i+1}$, and the process repeats.

GP ODEs for Tractography. For the DTI shortest path problem (1), we make the following specific choices: For the prior covariance we use the Kronecker form of (4), with a Gaussian kernel $k(t_i, t_j) = \exp\left[-(t_i - t_j)^2/(2\lambda^2)\right]$, which has one free model parameter, the length scale λ. In contrast to [12], we estimate both V and λ with a marginal moment matching method instead of evidence maximization [17, §2.3], which allows for more adaptive solutions. Like [12], we include boundary conditions analytically in the prior belief as direct observations of c with vanishing noise, using a regular grid with $2N$ observations spread over the domain $t = [0, 1]$. We initialize the prior mean for the solver from a discrete shortest path $v_0 = a, v_1, \ldots, v_N = b$, which we pre-process with a GP least-squares smoother (with a square exponential kernel function, whose parameters are optimized using evidence maximization [17, §5.3] separately for each output dimension). This gives a smooth prior mean function for the GP solver.

GP ODE Solutions. The output of the ODE solver is a posterior GP belief $\mathcal{GP}(c; \tilde{\mu}, \tilde{k})$. The repeated linear extrapolation to construct the $\tilde{c}(t_i)$ in the GP solver is structurally very similar to Runge-Kutta methods [10], but does not afford the strong analytic guarantees of those classic solvers. However, this algorithm, at any point during its runtime, retains a joint, consistent probability distribution over one locally distinct solution c and all its derivatives (cf. Fig. 1), from which *joint* samples (candidate solutions for the ODE) can be drawn. In contrast, classic solvers only return a single joint point estimate for c, with *local* error estimates that do not allow joint sampling.

3 Experiments

Data. Tractography was performed on pre-processed diffusion data of 40 subjects provided as a subsample from the Q3 release of the Human Connectome Project (HCP) [6, 7, 9, 20]. The pre-processed HCP diffusion data contains 270 diffusion directions distributed equally over 3 shells with b-values $= 1000, 2000$ and 3000 s/mm^2 [20]. DTI tensors for each voxel were computed with dtifit [1]. Segmentation was performed with FAST [22]. The cortico-spinal tract (CST) used for experiments was obtained from the expert annotated Catani tract atlas [3]. ROI atlases were constructed in "template space" by overlapping the tract atlas with regions defined by the Harvard-Oxford cortical and sub-cortical atlas [5] using the overlap with the brainstem, the hippocampus and the amygdala for one region, and the overlap with the superior frontal gyrus, the precentral gyrus and the postcentral gyrus for the second region. The warp fields provided by HCP were used to warp the ROIs from "template space" to "subject spaces".

Evaluation. We subsample 250 pairs of points at random from the ROIs for all 40 subjects and compute the discrete and continuous shortest paths. To compare the quality of the solutions we compute the set of voxels each path passes through. Taking the Catani atlas as a reference we measure *accuracy* defined as the percentage of voxels which are classified as belonging to the CST

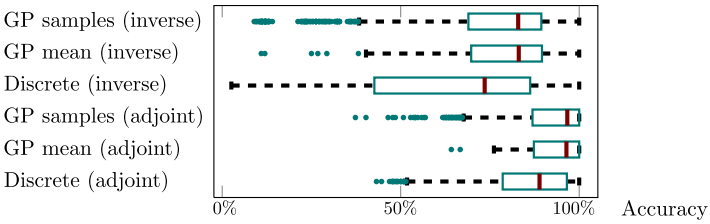


Fig. 2. Box plot of the evaluation score of different methods under different metrics

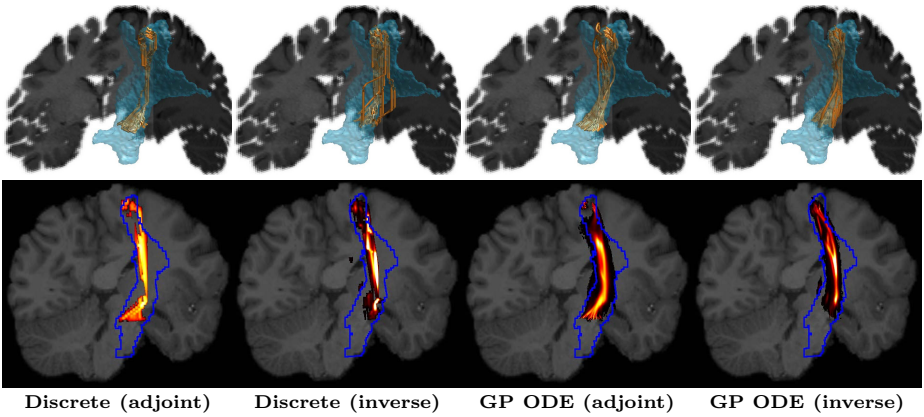


Fig. 3. Top: Geodesics under the inverse [13,14] and the adjoint metrics [8] in the right CST. Blue area shows voxels in the Catani atlas which at least one expert considered to be part of the tract. By considering different endpoints, bifurcating tracts can be discovered (1st and 3rd figure). **Bottom:** Density of discrete (left) and continuous (right) paths using two different metrics. CST of the Catani atlas as defined by at least one expert as reference (blue). **Also see the supplementary material[1].**

by at least one expert. We evaluate both the standard inverse metric [13,14] and the adjoint metric [8]. Fig. 2 shows the results for all 40 subjects.

Fig. 3 shows example paths from both algorithms and metrics computed on a single subject. The supplementary material[1] further contain an animation of samples from the GP solutions. Additionally, by counting the number of discrete paths going through each voxel, we produce a 3D density of paths. For each GP posterior we sample several curves, and proceed as in the discrete case. Fig. 3 shows a 2D heat map slice of the 3D density, chosen via principal component analysis to contain maximal variance.

4 Discussion and Conclusion

We use probabilistic numerics to compute shortest paths for tractography. This captures the uncertainty inherent in the shortest path computation which can

[1] `http://probabilistic-numerics.org/ODEs.html`

be used when visualizing the results. Sampling and averaging from the GP posterior allows us to marginalize over its uncertainty. We utilize this to generate heat maps of path densities. Compared to discrete methods, these heat maps appear smoother due to marginalization, but also sharper due to the continuous description of the sampled solutions.

Fig. 3 shows the classic "spaghetti plot" used for visualizing tractography results. The discrete solutions show tendencies to straight line segments connected by rather sharp turns which cannot solely be explained by the discretization error. In general, the continuous solutions bend less drastically, as we would expect from actual fibers. The figure only shows the mean function of the GP estimates of the geodesics; **the supplements contain an animation of the uncertainty**[1]. This provides a visualization of the solutions which makes it very clear that individual paths cannot, and *should not*, be interpreted as individual fibers in the brain — a common misinterpretation of "spaghetti plots".

We introduce a quality measure to compare different methods. First, we compare the standard inverse metric [13, 14] to the recently suggested adjoint metric [8]; empirical results show that the theoretically strong adjoint metric is consistently better. Secondly, we find that the GP posteriors agree with experts slightly more often than the discrete solutions. Generally, samples from the posterior score lower than the mean prediction, but this is to be expected since the Gaussian distribution puts non-zero probability mass everywhere.

A general disadvantage of shortest path tractography is that there will always be *some* path connecting any two points, whether it is anatomically there or not. This can be alleviated to some extent by discarding improbable paths.

Recent work by Wassermann et al. [21] illustrates that GPs form a particularly useful representation of shortest paths for population studies of brain connectivity. Our GP solution to the tractography problem lends itself particularly well to this type of population analysis as it avoids the post-hoc GP fitting used by Wassermann et al. This will be further investigated in future work.

Acknowledgements. Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. A.F. is funded in part by the Danish Council for Independent Research (DFF), Technology and Production Sciences. S.H. is funded in part by the Villum Foundation; the DFF, Natural Sciences; and an `amazon.com` machine learning in education award.

References

1. Basser, P.J., Mattiello, J., LeBihan, D.: Estimation of the effective self-diffusion tensor from the NMR spin echo. J. Magn. Reson. B 103(3), 247–254 (1994)
2. Basser, P.J., Pajevic, S., Pierpaoli, C., Duda, J., Aldroubi, A.: In vivo fiber tractography using DT-MRI data. Magnetic Resonance in Medicine 44(4), 625–632 (2000)
3. Catani, M., de Schotten, M.T.: A diffusion tensor imaging tractography atlas for virtual in vivo dissections. Cortex 44(8), 1105–1132 (2008)

4. Chkrebtii, O., Campbell, D., Girolami, M., Calderhead, B.: Bayesian uncertainty quantification for differential equations. arXiv stat.ME 1306.2365 (June 2013)
5. Desikan, R.S., et al.: An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31(3), 968–980 (2006)
6. van Essen, D., et al.: The Human Connectome Project: a data acquisition perspective. NeuroImage 62(4), 2222–2231 (2012)
7. van Essen, D., et al.: The WU-Minn Human Connectome Project: an overview. NeuroImage 80, 62–79 (2013)
8. Fuster, A., et al.: A novel Riemannian metric for geodesic tractography in DTI. In: Computational Diffusion MRI and Brain Connectivity. Mathematics and Visualization, pp. 97–104 (2014)
9. Glasser, M., et al.: The minimal preprocessing pipelines for the human connectome project. NeuroImage 80, 105–124 (2013)
10. Hairer, E., Nørsett, S., Wanner, G.: Solving Ordinary Differential Equations I – Nonstiff Problems. Springer (1987)
11. Hao, X., Whitaker, R.T., Fletcher, P.T.: Adaptive Riemannian metrics for improved geodesic tracking of white matter. In: Székely, G., Hahn, H.K. (eds.) IPMI 2011. LNCS, vol. 6801, pp. 13–24. Springer, Heidelberg (2011)
12. Hennig, P., Hauberg, S.: Probabilistic solutions to differential equations and their application to Riemannian statistics. In: AISTATS. JMLR, W&CP, vol. 33 (2014)
13. Lenglet, C., Deriche, R., Faugeras, O.: Inferring white matter geometry from diffusion tensor MRI: Application to connectivity mapping. In: Pajdla, T., Matas, J.(G.) (eds.) ECCV 2004. LNCS, vol. 3024, pp. 127–140. Springer, Heidelberg (2004)
14. O'Donnell, L., Haker, S., Westin, C.-F.: New approaches to estimation of white matter connectivity in diffusion tensor MRI: Elliptic PDEs and geodesics in a tensor-warped space. In: Dohi, T., Kikinis, R. (eds.) MICCAI 2002, Part I. LNCS, vol. 2488, pp. 459–466. Springer, Heidelberg (2002)
15. Parker, G., et al.: A framework for a streamline-based probabilistic index of connectivity (PICo) using a structural interpretation of MRI diffusion measurements. J. Magn. Reson. Imaging 18(2), 242–254 (2003)
16. Prados, E., et al.: Control theory and fast marching techniques for brain connectivity mapping. In: CVPR, pp. 1076–1083 (2006)
17. Rasmussen, C., Williams, C.: GPs for Machine Learning. MIT Press (2006)
18. Schultz, T., et al.: Fuzzy fibers: Uncertainty in dMRI tractography. arXiv cs.CV 1307.3271 (2013)
19. Skilling, J.: Bayesian solution of ordinary differential equations. Maximum Entropy and Bayesian Methods, Seattle (1991)
20. Sotiropoulos, S.N., et al.: Effects of image reconstruction on fiber orientation mapping from multichannel diffusion MRI: reducing the noise floor using SENSE. Magn. Reson. Med. 70(6), 1682–1689 (2013)
21. Wassermann, D., Rathi, Y., Bouix, S., Kubicki, M., Kikinis, R., Shenton, M., Westin, C.-F.: White matter bundle registration and population analysis based on gaussian processes. In: Székely, G., Hahn, H.K. (eds.) IPMI 2011. LNCS, vol. 6801, pp. 320–332. Springer, Heidelberg (2011)
22. Zhang, Y., et al.: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE TMI 20(1), 45–57 (2001)

Construct and Assess Multimodal Mouse Brain Connectomes via Joint Modeling of Multi-scale DTI and Neuron Tracer Data

Hanbo Chen[1], Yu Zhao[1], Tuo Zhang[1,2], Hongmiao Zhang[3], Hui Kuang[3], Meng Li[3], Joe Z. Tsien[3], and Tianming Liu[1]

[1] Cortical Architecture Imaging and Discovery Laboratory, Department of Computer Science and Bioimaging Research Center, The University of Georgia, Athens, GA, USA
{cojoc,zhaoyu,ztlxl,tliu}@uga.edu
[2] School of Automation, Northwestern Polytechnical University, Xi'an, China
[3] Brain and Behavior Discovery Institute,
Medical College of Georgia at Georgia Regents University, GA, USA
{HOZHANG,HKUANG,MLI,JTSIEN}@gru.edu

Abstract. Mapping the neuronal wiring diagrams in the brain at multiple spatial scales has been one of the major brain mapping objectives. Macro-scale medical imaging modalities such as diffusion tensor imaging (DTI) and meso-scale biological imaging such as serial two-photon tomography have emerged as the prominent tools to reveal structural connectivity patterns at multiple scales. However, a significant gap that whether/how DTI data and microscopic data are correlated with each other for the same species of mammalian brains, e.g., mouse brains, has been rarely explored. To bridge this knowledge gap, this work aims to construct multi-modal mouse brain connectomes via joint modeling of macro-scale DTI data and meso-scale neuronal tracing data. Specifically, the high-resolution DTI data and its streamline tractography result are mapped to the Allen Mouse Brain Atlas, in which the high-density axonal projections were already mapped by microscopic serial two-photon tomography. Then, multi-modal connectomes were constructed and the multi-view spectral clustering method is employed to assess consistent and discrepant connectivity patterns across the multi-scale multi-modal connectomes. Experimental results demonstrated the importance of fusing multimodal, multi-scale imaging modalities for structural connectivity and connectome mapping.

Keywords: Multi-scale connectome, DTI, neuron tracer, brain mapping.

1 Introduction

Creating structural wiring maps of the brain (or named brain connectome) is believed critical in understanding the mind [1, 2]. With the emerging biomedical imaging techniques, mapping structural connection wiring patterns at different scales with relatively high resolutions is made possible. For instance, on the macro-scale, with noninvasive diffusion tensor imaging (DTI), long-range fiber connections are being studied widely, e.g., the ongoing Human Connectome Project (HCP) [3]. On the meso-scale, Allen

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 273–280, 2014.

Mouse Brain Connectivity Atlas released the first meso-scale mouse brain connectivity atlas based on neuron tracer and serial two-photon (STP) tomography recently[4]. With these new neuroimaging methods, our understanding of the brain has been significantly advanced at different scales, e.g., the ongoing Brain Decoding Project (http://braindecodingproject.org) is trying to merge data in different scales to unveil secrete of memory.

Essentially, multimodal neuroimaging data across spatial scales tend to provide rich complementary information that forms a comprehensive picture to understand the same biological question [1]. At the same time, each neuroimaging modality has its own limitations in terms of resolutions, quantification, noises, and variability. It is likely that the limitations in one modality could be made up by another. For instance, DTI is good at mapping global patterns of long-range brain wiring diagrams, but has been known to be limited in resolving crossing fibers, sensitive to noises and artifacts, possessing limited spatial resolution and etc. In comparison, STP tomography has very high spatial resolution and can trace complex axonal connection patterns, but is limited in capturing global wiring diagrams (which entails many injection sites) and it is very costly and time-consuming. Therefore, it would be much desirable to fuse multimodal, multi-scale imaging modalities to gain comprehensive and robust structural connectivity and connectome mapping.

Fortunately, with the publicly available data sources of both high-resolution DTI data and STP tomography data of mouse brains, both of which can be mapped to the same reference atlas, it becomes feasible and desirable to build multi-modal mouse brain connectomes via joint modeling of macro-scale DTI data and meso-scale neuronal tracing data and to assess consistent and discrepant connectivity patterns across these connectomes. Specifically, we defined the connectome nodes based on the fine-granularity neuroanatomic regions in the Allen Mouse Brain Atlas (ABA) and the connectivity edges based on the connection strengths derived from the DTI or STP tomography data. Then we employed a multi-view spectral clustering algorithm to group the connectomes into sub-networks, based on which agreements and discrepancies between multi-modal connectomes, and interesting results were obtained.

2 Method

2.1 Experimental Materials

Axonal projection images were downloaded from publicly available Allen Mouse Brain Connectivity Atlas (ACA) (http://connectivity.brain-map.org/). Images obtained from neuron tracing experiments covering the whole brain of mouse were applied in this study. In each experiment, rAAV tracer was injected to certain anatomical region of a mouse brain to label the projection from this region to the whole brain (Fig. 1(d)). After fixation and dissection, the mouse brain was then sliced (100 μm thick) and high resolution image (0.35 μm/pixel) was obtained for each slice with STP tomography [5]. Then the images were processed with injection site manually annotated by experts and a 3-D image stack was obtained and registered to the 3-D reference atlas space for analysis. It should be noted that, for the purpose of efficiency, all the 1378 injection sites were selected on the right hemisphere of mouse brain.

Meanwhile, high-resolution DTI data of an adult mouse brain was downloaded from the publicly available Mouse BIRN Data Repository and applied for this study. The data acquisition parameters and preprocessing pipeline were detailed in [6]. DTI data was aligned to the 3-D reference atlas in Allen Mouse Brain Atlas (ABA) via FSL FLIRT [7] (Fig. 1(c)). Streamline fiber tractography was performed using DTI Studio [8] and cortical surface was reconstructed for the purpose of visualization [9].

To analyze large-scale mouse brain connectomes in different modalities, we adopted the annotation of mouse brain's anatomical structure downloaded from ABA (http://mouse.brain-map.org/). As shown in Fig. 1(a), these brain regions were manually annotated by experts in the 3-D reference atlas of ABA. Notably, the original annotation file did not differentiate brain regions by hemisphere. To analyze cross hemisphere connections, we manually selected 159 brain regions and separated them by hemisphere, resulting in 3 groups of brain regions: left hemisphere, right hemisphere, and cross hemisphere. Finally, 471 regions of interest (ROIs) were applied for the construction and analysis of mouse brain connectome, as shown in Fig. 1(b).

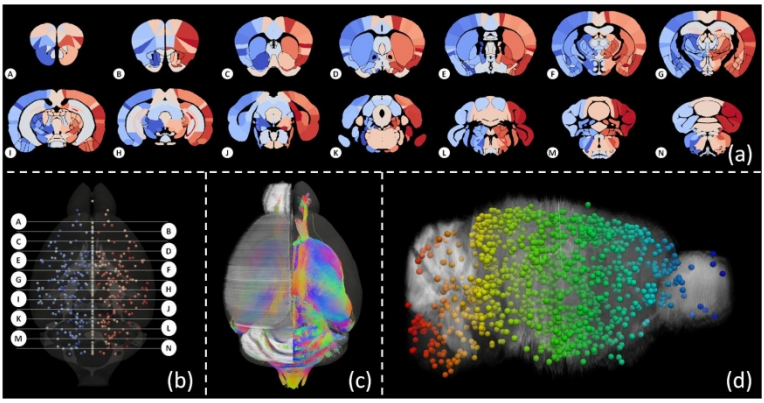


Fig. 1. (a) Fine-granularity mouse brain region annotation in ABA. 14 slices were selected for visualization. (b) Centers of all the regions we used in our analysis. Annotated regions are color-coded such that regions in the left hemisphere are blue while regions in the right hemisphere are red. (c) Joint visualization of 3-D reference atlas (left) and registered DTI-derived fiber tracts (right). (d) Visualization of injection sites in the ACA.

2.2 Cross-Validation of DTI-Derived Tracts and Neuron Traces

To quantitatively compare DTI-derived tracts and neuron traces, we adopted the Hausdorff distance [10] to measure the distance or discrepancy between streamlines of DTI-derived tracts and the neuron tracers, which were all aligned into the ABA space. Specifically, given a streamline $\boldsymbol{F}$ represented by a set of connected vertices and a trace $\boldsymbol{T}$ represented by a set of voxels, for each voxels in $\boldsymbol{T}$, its shortest distance to $\boldsymbol{F}$ will be calculated; and the Hausdorff distance is defined as the largest of the shortest distances to $\boldsymbol{F}$ of all voxels in $\boldsymbol{T}$:

$$Hausdorff(F,T) = \max(\{D \mid D_i = \min(\{\|V_i - P_k\| \mid P_k \in F\}), V_i \in T\}) \quad (1)$$

where V_i and P_k are the 3D coordinates of trace voxels and fiber vertices.

2.3 Construct Mouse Brain Connectomes

To compare large-scale mouse brain connectomes in different DTI and STP tomography modalities, the connections between ROIs introduced in section 2.1 were computed. As for DTI-based tracts, the fiber streamlines going through the brain volume of each ROI were extracted as the projection from/to the ROI. The number of fibers connecting each pair of ROIs was then counted as the connection strength between ROIs. To facilitate the quantitative analysis, the connection strength has been normalized to 1 such that values larger than 1000 will be set to 1 and the rest will be divided by 1000. The resulted connection matrix is shown in Fig. 2(a).

As for neuron tracer data, for each experiment, the density of projection in each ROI was applied to measure the connection strength. As rAVV tracer traces neurons anterogradely, for each tracing experiment, a vector has been generated indicating the connection strength from the injection to all the ROIs. And the connectivity matrix was constructed such that i^{th} row is obtained by averaging the vectors corresponding to experiments with neuron tracer injected to ROI i. However, as all the injection sites were selected on the right hemisphere of the brain, the connection from the left hemisphere to the rest of the brain is largely missing (Fig. 2 (b)). By assuming that connectome in mouse brain is symmetric (largely true based on the observation of DTI data and the known symmetry of mouse brains), we mapped the connection from the left hemisphere to the right hemisphere (Fig. 2 (c)).

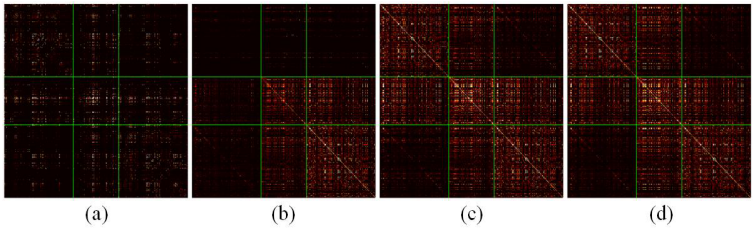


Fig. 2. Mouse brain's connectivity matrices. Each matrix was segmented by brain regions: from left to right, top to bottom are left hemisphere ROIs, across hemisphere ROIs, right hemisphere ROIs, respectively. Contrast has been enhanced for the purpose of visualization. (a) DTI derived brain connectome. (b)-(d) Neuron tracer derived brain connectomes. (b) Original connectivity matrix. (c) Symmetrically mapped connectome. (d) Symmetrified connectome.

2.4 Assess Mouse Brain Connectomes

The normalized-cut spectral clustering was applied to obtain sub-networks of the brain [11]. Basically, the salient eigenvectors of the connectivity matrix was applied to bi-partition the graph iteratively and the number of clusters will be determined by the threshold set for normalized cut. However, to solve the eigenvector problem, a matrix needs to be symmetric, which is not true for the connectivity matrix based on neuron tracer. Thus, we symmetrify the matrix by adding it to its transpose. As shown in Fig. 2(d), the symmetrified connectivity matrix does not change much by visual inspection. Meanwhile, a multi-view spectral clustering procedure has also been

applied to infer common clusters for these two modalities [11]. Specifically, the connectivity matrix will be projected to the eigenspace of the matrix of the other modality to maximize the agreement between them. By doing so, the common part of two connectivity matrices will be enhanced, while the discrepant part will be weakened. Also, the common clusters of two could be achieved.

3 Result

3.1 Comparison of DTI-Based Tracts and Neuron Traces

We calculated the Hausdorff distances between forty two thousands DTI-based tracts obtained by streamline tractography and 1378 neuron traces in the same ABA space. All of the DTI-based tracts are visualized in Fig. 3(a) and are color-coded by their minimum Hausdorff distance to the neuron traces. As all of the injection sites were picked on the right hemisphere, DTI-based tracts in the left hemisphere have longer Hausdorff distance in comparison with those in the right hemisphere. On average, the Hausdorff distance for the DTI-based tracts that projected from/to the right hemisphere is 0.68±0.43 mm, which is relatively small considering the size of mouse brain. Considering possible misalignment and false elimination of trace foreground with selected threshold, we assumed when Hausdorff distance is less than 1mm, the DTI-based tract has correspondence to the neuron trace. Two neuron tracer experiments and the corresponding DTI-based tracts were visualized in Fig. 3(b)-(c). It can be seen that DTI-based tracts truly possess correspondence to the real neuronal projections, *which validates that fibers derived from DTI offer reliable information of axonal connection and could be applied in the study of brain connectome*.

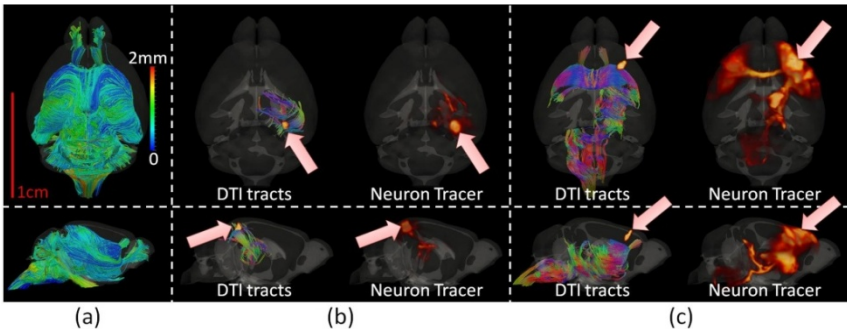


Fig. 3. (a) Visualization of all DTI-based fiber tracts. Tracts are color-coded by their minimum Hausdorff distances to neuron traces. (b)-(c) Visualization of two examples of neuron tracer experiments and the corresponding DTI-based tracts. Locations of injection site of neuron tracer in each experiment are highlighted by pink arrows.

Meanwhile, we noticed that when computing large-scale whole brain connectomes via DTI-based tracts, some brain regions are rarely connected. These regions were shown in Fig. 4(a). This might be partially because DTI could only detect axonal fiber

when big bundle of axons orient in the same direction (e.g. in white matter). As white matter volume is relatively small in mouse brain, when some ROIs do not have direct connection with white matter due to their nature or the possible misalignment between DTI data and 3-D reference atlas, connections cannot be identified by DTI-based tracts for these ROIs. We then obtained the common connections and unique connections obtained by two multi-scale modalities (Fig. 4(b)-(d)). It can be seen in Fig. 4(b) that those two multi-scale modalities inferred a large number of common connections, which to some extent cross-validates each other. However, by calculating the average Euclidean distance between connected regions, we found that the connection lengths of common connections (2.76 mm) are relatively shorter than unique connections. Meanwhile, connections identified by DTI only are much longer (4.4 mm) in comparison. This result suggests that both multi-scale modalities work well in identifying short connections. Also, the long-range connections generated by DTI tractography may be not reliable and some may be missing. However, as the neuron tracer signal will degenerate when travelling long distance and the signal strength will be impacted by the dose of tracers injected, it is equally possible that some long distance connections are missed due to the weak signal in neuron tracer experiments. Thus, it will be very important to compare the results in different modalities to achieve better understanding of the structural wiring diagram of the brain.

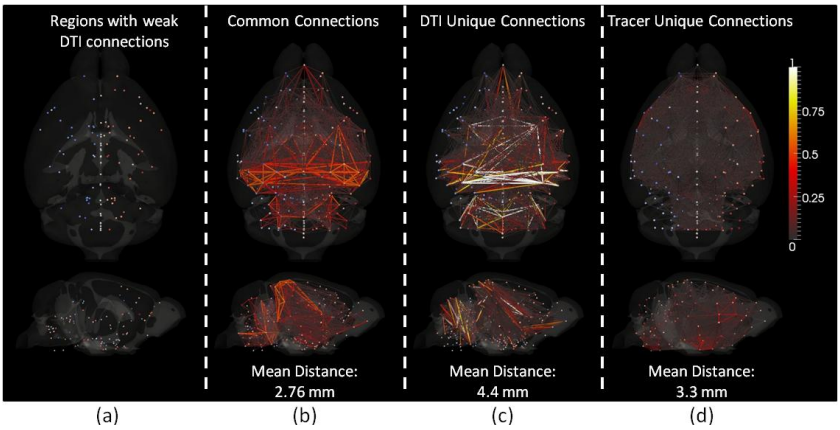


Fig. 4. (a) Visualization of brain locations of regions that have weak DTI-derived connections. (b)-(d) Visualization of common and unique connections between DTI and neuron tracer. Connections are color-coded and scaled by connection strength accordingly (color bar on right).

3.2 Joint Assessment of Multimodal Mouse Brain Connectomes

The multi-scale brain connectomes derived from different modalities were firstly assessed by graphical measurements. Notably, when analyzing DTI-derived brain connectome, we excluded those ROIs that have no connections as they can cause error in some measurements. As shown in Table 1, fewer connections have been identified by DTI-based tracts, which results in a smaller average degree and density. To measure the graph topology, the average clustering coefficients and average path

length were computed and compared with the corresponding randomized graphs [12]. It is evident that the overall graph topological structures of these two connectomes by different modalities are similar and both of them are small-world networks.

Table 1. Graph-theoretic measurements of multimodal connectomes

	average degree	density	average clustering coefficient			average path length		
			original	randomized	o/r	original	randomized	o/r
DTI	88	0.25	0.71	0.32	2.18	1.83	1.67	1.09
tracer	137	0.3	0.69	0.32	2.13	1.74	1.68	1.04

We then compared the sub-networks obtained by spectral clustering algorithm. The results are shown in Fig. 5. For those ROIs with no connection by DTI-derived tracts, we manually labeled them as a class when performing clustering based on DTI-derived connectivity matrix or jointly with both matrices. As highlighted by yellow and blue arrows, after clustering, some common clusters of two modalities were clearly identified. As highlighted by red arrows, both modalities separated cerebellum from other parts of the brain. By using multi-view spectral clustering, all the common clusters inferred by two modalities could be recognized. More analysis will be conducted to further understand these clusters.

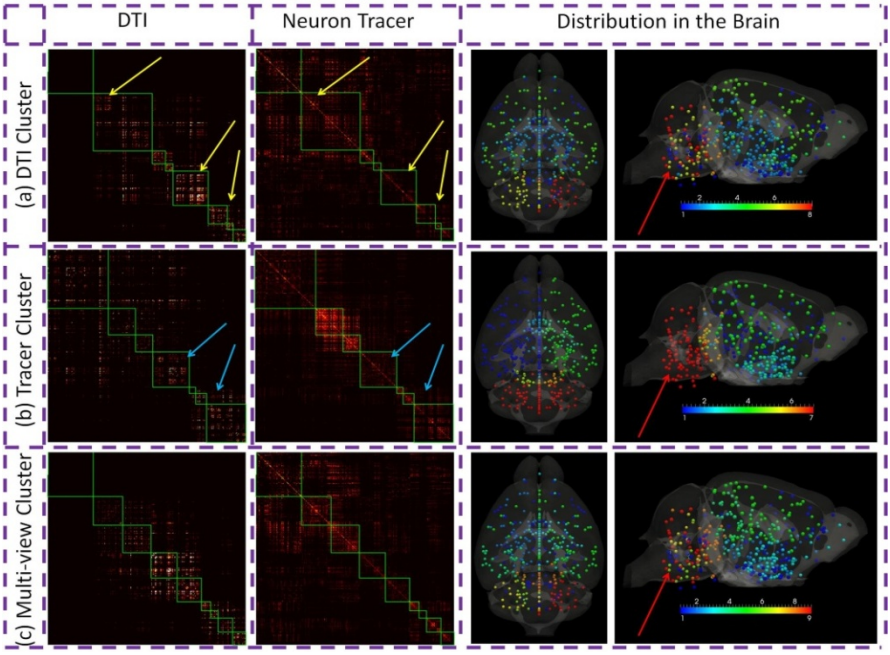


Fig. 5. Connectivity matrices based on DTI-based tracts (left column) and neuron tracers (middle column) were re-arranged by clusters with each cluster highlighted by green box accordingly. Clusters were visualized in the mouse brain with each cluster color-coded by certain color (right column). (a) Clusters obtained based on DTI-based tracts; (b) Clusters obtained based on neuron tracers; (c) Clusters obtained by both with multi-view spectral clustering.

4 Discussion and Conclusion

This paper presents a novel computational framework to construct multi-modal brain connectomes via joint modeling of macro-scale DTI data and meso-scale neuronal tracing data. The publicly available data sources of DTI data and STP tomography data of mouse brains were mapped to the same Allen mouse brain atlas for the construction of multi-modal connectomes. Then consistent and discrepant connectivity patterns across these connectomes, as well as the graph theoretic measurements, were assessed. Our results quantitatively demonstrated that both DTI and neurnal tracing are valuable complementary tools for structural connectivity and connectome mapping, and particularly the fusion of these multimodal, multi-scale imaging modalities has a potential to create synergy to facilitate connectome-related research.

Acknowledgement. Data used for this study were downloaded from the Mouse BIRN Data Repository, supported by grants to the Mouse BIRN (U24-RR021760) Testbed funded by the National Center for Research Resources at the National Institutes of Health, U.S.A.

References

1. He, B., Coleman, T., Genin, G.M., Glover, G., Hu, X., Johnson, N., Liu, T., Makeig, S., Sajda, P., Ye, K.: Grand challenges in mapping the human brain: NSF workshop report. IEEE Trans. Biomed. Eng. 60, 2983–2992 (2013)
2. ACD Interim Report: Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Working Group (2013)
3. Sotiropoulos, S.N., Jbabdi, S., Xu, J., Andersson, J.L., Moeller, S., Auerbach, E.J., Glasser, M.F., Hernandez, M., Sapiro, G., Jenkinson, M., Feinberg, D.A., Yacoub, E., Lenglet, C., Van Essen, D.C., Ugurbil, K., Behrens, T.E.J.: Advances in diffusion MRI acquisition and processing in the Human Connectome Project. Neuroimage 80, 125–143 (2013)
4. Oh, S.W., et al.: A mesoscale connectome of the mouse brain. Nature 508, 207–214 (2014)
5. Ragan, T., Kadiri, L.R., Venkataraju, K.U., Bahlmann, K., Sutin, J., Taranda, J., Arganda-Carreras, I., Kim, Y., Seung, H.S., Osten, P.: Serial two-photon tomography for automated ex vivo mouse brain imaging. Nat. Methods 9, 255–258 (2012)
6. Zhang, J., van Zijl, P.C.M., Mori, S.: Three-dimensional diffusion tensor magnetic resonance microimaging of adult mouse brain and hippocampus. Neuroimage 15, 892–901 (2002)
7. Jenkinson, M., Smith, S.: A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5, 143–156 (2001)
8. Jiang, H., van Zijl, P.C.M., Kim, J., Pearlson, G.D., Mori, S.: DtiStudio: resource program for diffusion tensor computation and fiber bundle tracking. Comput. Methods Programs Biomed. 81, 106–116 (2006)
9. Liu, T., Nie, J., Tarokh, A., Guo, L., Wong, S.T.C.: Reconstruction of central cortical surface from brain MRI images: method and application. Neuroimage 40, 991–1002 (2008)
10. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the Hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. 15, 850–863 (1993)
11. Chen, H., Li, K., Zhu, D., Jiang, X., Yuan, Y., Lv, P., Zhang, T., Guo, L., Shen, D., Liu, T.: Inferring group-wise consistent multimodal brain networks via multi-view spectral clustering. IEEE Trans. Med. Imaging. 32, 1576–1586 (2013)
12. Bullmore, E., Sporns, O.: Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Rev. Neurosci. 10, 186–198 (2009)

Designing Single- and Multiple-Shell Sampling Schemes for Diffusion MRI Using Spherical Code

Jian Cheng, Dinggang Shen, and Pew-Thian Yap

Department of Radiology and BRIC, The University of North Carolina at Chapel Hill, USA
{jian_cheng,dgshen,ptyap}@med.unc.edu

Abstract. In diffusion MRI (dMRI), determining an appropriate sampling scheme is crucial for acquiring the maximal amount of information for data reconstruction and analysis using the minimal amount of time. For single-shell acquisition, uniform sampling without directional preference is usually favored. To achieve this, a commonly used approach is the Electrostatic Energy Minimization (EEM) method introduced in dMRI by Jones et al. However, the electrostatic energy formulation in EEM is not *directly* related to the goal of optimal sampling-scheme design, i.e., achieving large angular separation between sampling points. A mathematically more natural approach is to consider the Spherical Code (SC) formulation, which aims to achieve uniform sampling by maximizing the minimal angular difference between sampling points on the unit sphere. Although SC is well studied in the mathematical literature, its current formulation is limited to a single shell and is not applicable to multiple shells. Moreover, SC, or more precisely continuous SC (CSC), currently can only be applied on the *continuous* unit sphere and hence cannot be used in situations where one or several subsets of sampling points need to be determined from an existing sampling scheme. In this case, discrete SC (DSC) is required. In this paper, we propose novel DSC and CSC methods for designing uniform single-/multi-shell sampling schemes. The DSC and CSC formulations are solved respectively by Mixed Integer Linear Programming (MILP) and a gradient descent approach. A fast greedy incremental solution is also provided for both DSC and CSC. To our knowledge, this is the first work to use SC formulation for designing sampling schemes in dMRI. Experimental results indicate that our methods obtain larger angular separation and better rotational invariance than the generalized EEM (gEEM) method currently used in the Human Connectome Project (HCP).

1 Introduction

Diffusion MRI (dMRI) is a unique technique for exploring the underlying tissue properties of white matter in the human brain. A central problem in dMRI is to reconstruct the MR signal attenuation $E(\mathbf{q})$ from a limited number of measurements in the $\mathbf{q}$-space and to estimate some meaningful quantities such as the Ensemble Average Propagator (EAP) and the Orientation Distribution Function (ODF). An effective $\mathbf{q}$-space sampling scheme is critical for the acquisition of maximum information with minimum time cost. Since white matter fascicles traverse the brain in a wide range of directions, a uniform sampling scheme with no directional preference is often preferred [1].

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 281–288, 2014.

In the last decade, two approaches have been widely used for designing sampling schemes for single-shell acquisition. The first approach involves tessellation of a unit sphere using basic shapes such as the icosahedron. Using such a spherical tessellation approach, however, one is unable to generate a sampling scheme with an arbitrary number of sample points. The second approach is the Electrostatic Energy Minimization (EEM) method, which was introduced in dMRI by Jones et al. [1]. Some best known solutions to the EEM problem have been collected in CAMINO [2]. EEM was also recently generalized for multi-shell sampling [3,4] with staggered samples in different shells and has been adopted in the Human Connectome Project (HCP) [5]. However, the electrostatic energy formulation in EEM is not directly related to the goal of sampling scheme design, which is to maximize the angular separation between sampling points, and it is still unknown why electrostatic energy matters in dMRI reconstruction.

A good sampling scheme should have large angular separation such that the reconstruction has large angular resolution and good rotational invariance. Thus a mathematically more natural way for sampling scheme design is to maximize the minimal angular difference between sampling points, i.e., *covering radius*, on a unit sphere. Determining such point configuration is essentially the Spherical Code (SC) problem[1], and there are a collection of best known solutions for the SC problem in $\mathbb{S}^2$ [6][2]. Although SC is well studied in the mathematical literature, its current formulation is limited to a single shell and is not applicable to multiple shells. Moreover, SC, or more precisely continuous SC (CSC), currently can only be applied on the *continuous* unit sphere and hence cannot be used in situations where several subsets of sampling points need to be determined from an existing sampling scheme. In this case, discrete SC (DSC) is required, where the solution space is discrete and determined by a set of predetermined sampling points.

In this paper, we propose novel CSC and DSC methods for designing single-/multi-shell sampling schemes. We propose a greedy incremental estimation for rapid generation of solutions to the DSC and CSC problems, a Mixed Integer Linear Programming (MILP) method to solve the DSC problem, and a Riemannian gradient descent method to solve the CSC problem. Experimental results indicate that the proposed methods are capable of yielding larger covering radius and better rotational invariance than the state-of-the-art generalized EEM (gEEM) method currently used in the HCP [4,5].

2 Designing Sampling Scheme Using Spherical Code

2.1 Discrete Spherical Code (DSC) and Continuous Spherical Code (CSC)

For single-shell sampling, the SC problem is to determine a set of K points $\{\mathbf{u}_i\}_{i=1}^K$ such that the minimal distance between these points is maximized, i.e.,

$$\max_{\{\mathbf{u}_i \in \mathbb{D}\}_{i=1}^K} d(\{\mathbf{u}_i\}_{i=1}^K), \qquad d(\{\mathbf{u}_i\}_{i=1}^K) = \min_{i \neq j} \arccos |\mathbf{u}_i^T \mathbf{u}_j|, \tag{1}$$

where $d(\{\mathbf{u}_i\}_{i=1}^K)$ is the minimal angular distance, or called *covering radius*, of point set $\{\mathbf{u}_i\}_{i=1}^K$, and $\mathbb{D} \subseteq \mathbb{S}^2$ is the solution domain. If $\mathbb{D} = \mathbb{S}^2$, Eq. (1) is a CSC problem

[1] http://mathworld.wolfram.com/SphericalCode.html
[2] http://neilsloane.com/grass/dim3/

for selecting K points from throughout the unit sphere $\mathbb{S}^2$. If $\mathbb{D} = \{\mathbf{u}_n\}_{n=1}^N$, a set of N predetermined points on $\mathbb{S}^2$, then Eq. (1) is a DSC problem for selecting K from N points. We use the absolute value of $\mathbf{u}_i^T\mathbf{u}_j$ in Eq. (1) because antipodal symmetric samples have the same role in diffusion MRI data reconstruction. Note that the original SC in mathematics only means CSC, while in this paper it is the first time that we propose both CSC and DSC and generalize them for multi-shell case for designing sampling scheme in dMRI.

For multi-shell sampling, the SC problem is to find a set of points $\{\mathbf{u}_{s,i}\}$ by solving

$$\max_{\{\mathbf{u}_{s,i}\in\mathbb{D}\}} wS^{-1}\sum_{s=1}^{S} d(\{\mathbf{u}_{s,i}\}_{i=1}^{K_s}) + (1-w)d(\{\mathbf{u}_{s,i}\}_{i=1,\ldots,K_s;s=1,\ldots,S}), \tag{2}$$

where $\mathbf{u}_{s,i}$ is the i-th point on the s-th shell, S is the number of shells, K_s is the number of points on the s-th shell, and w is a weighting factor for balancing two terms. In Eq. (2), the first term is the mean covering radius of the S shells, and the second term is the covering radius for a combined shell containing all points from the S shells. Due to the second term, the estimated samples in different shell are staggered.

2.2 Greedy Incremental Solver

Similarly to EEM [7] and gEEM [3,4], we propose a greedy solver for incremental estimation of sampling schemes. Incremental estimation can be applied for both Eq. (1) and Eq. (2) when solving a DSC problem, i.e., when $\mathbb{D} = \{\mathbf{u}_i\}_{i=1}^N$. In step k, we estimate one point $\mathbf{u} \in \mathbb{D}$ that maximizes the cost function based on the $k-1$ points estimated in previous iterations. This incremental estimation technique can be applied to generate an approximate solution to a CSC problem, i.e., $\mathbb{D} = \mathbb{S}^2$, by approximating $\mathbb{S}^2$ using a large number of uniformly distributed points. In practice, 20481 points from a 7 order tessellation of the icosahedron are used. Incremental estimation can generate reasonable solutions in seconds.

2.3 DSC via Mixed Integer Linear Programming (MILP)

Instead of incrementally estimating samples one by one, Mixed Integer Linear Programming (MILP) can be used to estimate samples simultaneously. For $\mathbb{D} = \{\mathbf{u}_n\}_{n=1}^N$, Eq. (1) can be solved using MILP in Eq. (3a) as follows:

$$\max_{y,\{h_i\}_{i=1}^N} \; y \tag{3a}$$

$$\text{s.t. } \arccos(|\mathbf{u}_i^T\mathbf{u}_j|) \geq y - (2-h_i-h_j)M, \;\; \forall i > j \tag{3b}$$

$$d_{\text{LB}} \leq y \leq d_{\text{UB}}(2K) \tag{3c}$$

$$\sum_{i=1}^{N} h_i = K; \qquad h_i = 0,1, \; \forall i \tag{3d}$$

where $h_i = 1$ indicates that $\mathbf{u}_i$ is selected as one of the K points, d_{LB} is the lower bound of the covering radius, which can be set to 0 or the covering radius from an

existing sampling scheme, $d_{\text{UB}}(2K) = \arccos\sqrt{4-\csc^2\left(\frac{\pi K}{6(K-1)}\right)}$ is the theoretical upper bound of the covering radius for $2K$ points on $\mathbb{S}^2$ [8], and M is the difference between the maximal and minimal distances of any two points $\mathbf{u}_i, \mathbf{u}_j \in \mathbb{D}, i \neq j$. Note that 1) after solving MILP, the solution of y, denoted as y^*, is the covering radius of the selected K samples; 2) the constraint in Eq. (3b) only takes effect when $h_i = h_j = 1$, and is automatically satisfied when $h_i = 0$ or $h_j = 0$, because the chosen M is large enough such that $\arccos(|\mathbf{u}_i^T\mathbf{u}_j|) \geq 0 \geq y^* - (2 - h_i - h_j)M$ when $h_i + h_j \leq 1$. Similarly, Eq. (2) can be solved using MILP in Eq. (4a) as follows:

$$\max_{\{y_s\},\{h_{s,i}\}} wS^{-1}\sum_{s=1}^{S} y_s + (1-w)y_0 \tag{4a}$$

$$\text{s.t. } \arccos(|\mathbf{u}_i^T\mathbf{u}_j|) \geq y_s - (2 - h_{s,i} - h_{s,j})M, \ \forall s, i > j \tag{4b}$$

$$\arccos(|\mathbf{u}_i^T\mathbf{u}_j|) \geq y_0 - (2 - h_{s,i} - h_{s',j})M, \ \forall s, s', i > j \tag{4c}$$

$$d_{\text{LB},s} \leq y_s \leq d_{\text{UB}}(2K_s), \ \forall s; \quad d_{\text{LB},0} \leq y_0 \leq d_{\text{UB}}(2\sum_{i=1}^{S} K_s) \tag{4d}$$

$$\sum_{i=1}^{N} h_{s,i} = K_s, \ \forall s; \quad \sum_{s=1}^{S} h_{s,i} \leq 1, \ \forall i; \quad h_{s,i} = 0, 1, \ \forall i, s \tag{4e}$$

where $h_{s,i} = 1$ indicates that $\mathbf{u}_i$ is selected as one of the K_s points on the s-th shell, $d_{\text{LB},s}$ and $d_{\text{UB}}(2K_s)$ are the lower and upper bounds of the covering radius on the s-th shell, and $d_{\text{LB},0}$ and $d_{\text{UB}}(2\sum_{i=1}^{S} K_s)$ are the lower and upper bounds for the combined shell with all points. Constraints in Eq. (4b) and Eq. (4c) are respectively for the first and second terms in Eq. (4a). $\sum_{s=1}^{S} h_{s,i} \leq 1$ makes $\mathbf{u}_i$ to be selected at most one shell such that the estimated samples are staggered in different shells.

MILP problem can be solved using branch and bound method which iteratively solves the relaxed LP program. In our implementation, we solve Eq. (3a) and Eq. (4a) using GUROBI [9], which can obtain the global solution or at least a reasonable solution within minutes for DSC. In practice, we progressively increase the lower bound d_{LB} based on the solutions estimated in previous iterations to find a better feasible solution within 10 minutes which is good enough in experiments.

2.4 CSC via Riemannian Gradient Descent

When $\mathbb{D} = \mathbb{S}^2$, Eq. (1) can be solved using a Riemannian gradient descent method. For each iteration, we detect the pairs of points $\{(\mathbf{u}_{p_0}, \mathbf{u}_{p_1})\}$ whose angular differences are equal to the minimal angular difference computed from all point pairs. Noting that the Euclidean gradient of function $d(\{\mathbf{u}_j\}_{j=1}^K)$ is

$$\frac{\partial d(\{\mathbf{u}_j\}_{j=1}^K)}{\partial \mathbf{u}_i} = \begin{cases} -\sum_{p_1} \frac{1}{\sqrt{1-(\mathbf{u}_{p_0}^T\mathbf{u}_{p_1})^2}} \text{sign}(\mathbf{u}_{p_0}^T\mathbf{u}_{p_1})\mathbf{u}_{p_1} & \text{if } i = p_0 \\ -\sum_{p_0} \frac{1}{\sqrt{1-(\mathbf{u}_{p_0}^T\mathbf{u}_{p_1})^2}} \text{sign}(\mathbf{u}_{p_0}^T\mathbf{u}_{p_1})\mathbf{u}_{p_0} & \text{if } i = p_1 \\ 0 & \text{if } i \notin \{p_0, p_1\} \end{cases} \tag{5}$$

the Riemannian gradient for $\mathbf{u}_i$ can be computed as [10]

$$\nabla_{\mathbf{u}_i} d(\{\mathbf{u}_j\}) = \frac{\partial d(\{\mathbf{u}_j\})}{\partial \mathbf{u}_i} - \left(\mathbf{u}_i^T \frac{\partial d(\{\mathbf{u}_j\})}{\partial \mathbf{u}_i}\right)\mathbf{u}_i. \tag{6}$$

Algorithm 1. CSC via Riemannian Gradient Descent

Input: Initialization $\{\mathbf{u}_{s,i}\}_{i=1}^{K_s}$, $s = 1, ..., S$.
Output: Refined $\{\mathbf{u}_{s,i}\}_{i=1}^{K_s}$, $s = 1, ..., S$.
$k = 0$, $t_0 = 0.1$, determine the set of point pairs $P^{(k)} = \{(\mathbf{u}_{p_0^s}, \mathbf{u}_{p_1^s})\}_{s=0}^{S}$ that have minimal distances in $S + 1$ shells;
repeat
  Record: $P_0^{(k)} \leftarrow P^{(k)}$;
  repeat
    Calculate Riemannian gradient $\boldsymbol{v}_{s,i}, \forall s, i$, using point-pair set $P_0^{(k)}$ and Eq. (6) ;
    if $\forall s, i$, $\|\boldsymbol{v}_{s,i}\| \leq \epsilon$ **then** break;
    Choose step size $t \in (0, t_0]$ via inexact line search;
    $\mathbf{u}_{s,i}^{(k+1)} = \mathrm{Exp}_{\mathbf{u}_{s,i}^{(k)}}\left(t\frac{\boldsymbol{v}_{s,i}}{\max_{s,i}\{\|\boldsymbol{v}_{s,i}\|\}}\right), \forall s, i$;
    Detect updated set $P^{(k+1)}$ based on $\{\mathbf{u}_{s,i}^{(k+1)}\}$;
    if $P^{(k+1)} \not\subseteq P_0^{(k)}$ **then** $P_0^{(k)} \leftarrow P_0^{(k)} \cup P^{(k+1)}$;
  until $P^{(k+1)} \subseteq P_0^{(k)}$;
  $k \leftarrow k + 1$;
until *Cost function does not change*;

Then the gradient descent update is performed using

$$\mathbf{u}_i^{(k+1)} = \mathrm{Exp}_{\mathbf{u}_i^{(k)}}\left(t\frac{\nabla_{\mathbf{u}_i} d(\{\mathbf{u}_j\})}{\max_i\{\|\nabla_{\mathbf{u}_i} d(\{\mathbf{u}_j\})\|\}}\right), \quad \mathrm{Exp}_{\mathbf{u}}(\boldsymbol{v}) = \mathbf{u}\cos\|\boldsymbol{v}\| + \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}\sin\|\boldsymbol{v}\|, \quad (7)$$

where the largest norm of gradient vectors is used for normalization of all gradient vectors, $\mathrm{Exp}_{\mathbf{u}}(\boldsymbol{v})$ is the exponential map [10] that maps the gradient vector $\boldsymbol{v}$ from the tangent space of the unit sphere at $\mathbf{u}$ to the unit sphere itself. Note that Riemannian gradient descent is performed on all $\{\mathbf{u}_i\}_{i=1}^{K}$ simultaneously. For the multi-shell case, point pairs $\{(\mathbf{u}_{p_0^s}, \mathbf{u}_{p_1^s})\}$ with minimal distances are detected from all S shells and the combined shell with all points. Then similar gradient descent is performed, taking the gradient of Eq. (4a) as a summation of gradients from $S + 1$ shells.

One important issue of the proposed Riemannian gradient descent method is that, after each gradient descent, the set of point pairs $\{(\mathbf{u}_{p_0^s}, \mathbf{u}_{p_1^s})\}$ with minimal distances may change, and the cost function and its gradient, which depends on these pairs, may also change. To solve this issue, in each iteration, we compare the updated set of point pairs with the previous pair set before the one step gradient descent, and if the updated pair set has some new pairs which are not in the previous pair set, then we re-perform the gradient descent using the joint set of these two sets. See Algorithm 1 for implementation details. The algorithm is efficient and converges in seconds.

Since the optimization problem is highly non-convex, a good initialization is important for a good solution. Initialization can be set as a set of random points, the solution given by incremental estimation, or the solution given by MILP. Note that when using MILP for the initialization, solving the CSC becomes slow when a large number of points are involved. In practice, we apply MILP to 321 points obtained from spherical tessellation and use the MILP solution as the initialization of the gradient descent method for CSC.

Table 1. Covering radii of multi-shell sampling schemes with number of samples 28×3 and 90×3 generated by various methods. Note that gEEM results for 90×3 are not available in [4], and EEM results from CAMINO are individually for each single shell (the scheme with 270 samples is not available).

	Shell 1 (28)	Shell 2 (28)	Shell 3 (28)	Combined (28×3)
gEEM [4]	22.2°	22.2°	22.0°	13.2°
Incr. gEEM [4]	19.2°	19.7°	19.3°	4.7°
Incr. CSC ($N = 20481$)	21.3°	19.3°	21.1°	10.5°
MILP ($N = 321$)	23.8°	23.8°	24.3°	13.3°
Incr. CSC + Grad. Desc.	24.3°	23.5°	24.2°	10.9°
MILP + Grad. Desc.	**25.7**$^\circ$	**25.7**$^\circ$	**25.4**$^\circ$	**13.6**$^\circ$
EEM (CAMINO) [1,2]	25.7°	25.7°	25.7°	15.6°
	Shell 1 (90)	Shell 2 (90)	Shell 3 (90)	Combined (90×3)
Incr. gEEM [4]	10.8°	10.3°	10.5°	2.4°
Incr. CSC ($N = 20481$)	10.4°	9.7°	10.4°	4.6°
MILP ($N = 321$)	13.3°	13.5°	13.3°	**7.9**$^\circ$
Incr. CSC + Grad. Desc.	13.0°	13.5°	12.8°	5.2°
MILP + Grad. Desc.	**14.6**$^\circ$	**15.0**$^\circ$	**14.8**$^\circ$	7.5°
EEM (CAMINO) [1,2]	15.1°	15.1°	15.1°	-

3 Experiments

Separation of Sampling Schemes. We evaluated the effectiveness of the proposed multi-shell MILP-based DSC method in Eq. (4a) with $w = 1$, by gauging whether it can separate points on a set of samples into several subsets, keeping the points in each subset as uniform as possible. We used the `subsetpoints` program in CAMINO [1,2] for comparison, which performs the same task by using simulated annealing. For this evaluation, we randomly mixed two sets of uniform points, one set consisting of 81 points generated by spherical tessellation and the other set consisting of 60 points from CAMINO generated by EEM. Separation of these 141 points into subsets respectively with 81 and 60 samples should ideally give results that match the original uniform point sets. Our method, which uses MILP, gave results that exactly match the original point sets within 5 seconds. `subsetpoints` in CAMINO saves the result every hour when it runs. It gave two incorrect points in each subset after running for 2 hours, 7 incorrect points after 8 hours, and the correct result after 9 hours. Although the correct result had been obtained, the program continued to run for hours until the simulated annealing temperature was finally small enough.

Multi-shell Angular Separation. We evaluated the effectiveness of the proposed multi-shell DSC (MILP) and CSC (incremental estimation) methods in generating a three shell sampling scheme, each shell consisting of K sampling points. We tested two cases: $K = 28$ and $K = 90$. These two cases were used such that we could compare our results with those given by gEEM and incremental gEEM [4]; results for 28 points per shell were reported in [4] and results for 90 points per shell were utilized in the HCP. MILP was used to select $K \times 3$ points from 321 points given by spherical tessellation. In incremental CSC, the selection was carried out using 20481 uniformly distributed points. Gradient descent was then used to refine these results. In Table 1, the

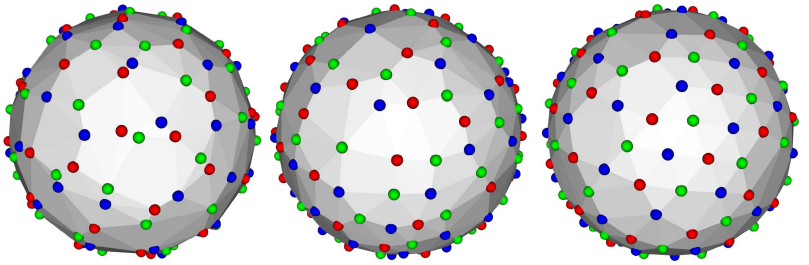


Fig. 1. Multi-shell sampling schemes with 28×3 samples generated by three methods and their combined covering radii showed in Table 1. The colors differentiate the sampling points from the three shells.

covering radii, i.e., the minimal angular differences, of these results were compared with those given by gEEM and incremental gEEM [4]. The multi-shell results of gEEM with $K = 28$ were extracted from [4]. The results of incremental gEEM with $K = 28, 90$, which have been used in HCP, were obtained from the website[3] created for [4]. The covering radii for single-shell results given by EEM in CAMINO are shown for reference. Fig. 1 visualizes the results with $K = 28$ generated by incremental gEEM, incremental CSC, and gradient descent with MILP initialization. Table 1 and Fig. 1 demonstrate clearly that the proposed MILP method and incremental CSC estimation yield larger covering radii than gEEM and incremental gEEM in all three shells and the combined shell containing all points. The proposed gradient descent method with MILP initialization yields the best angular separation in the multi-shell case, and its results are comparable with the single-shell results given by EEM in CAMINO, although the optimization was done with respect to all shells.

Rotational Invariance in Reconstruction. We tested the multi-shell sampling schemes with 28×3 samples in Table 1 on whether they give consistent reconstruction results for the synthetic signals generated by rotated models. A mixture of tensor model was used: $E(q\mathbf{u}) = 0.5\exp(-q^2\mathbf{u}^T\mathbf{D}_1\mathbf{u}) + 0.5\exp(-q^2\mathbf{u}^T\mathbf{D}_2\mathbf{u})$, where $b = q^2 = 1000, 2000, 3000\,\mathrm{s/mm^2}$, and the two tensors $\mathbf{D}_1$, $\mathbf{D}_2$ have the same eigenvalues $[1.7, 0.2, 0.2] \times 10^{-3}\,\mathrm{mm^2/s}$ and have a crossing angle of $60°$. This signal generation is repeated 20481 times by rotating the model according to 20481 uniformly distributed directions generated by spherical tessellation. We performed Spherical Polar Fourier Imaging (SPFI) with spherical order 6 and radial order 2 [11] to estimate the EAP profiles with radius of 15μm, detected the peaks of the EAP profiles, and compared the detected peaks with the ground-truth fiber directions in these 20481 tests. The means and standard deviations are shown in Table 2. Note that we have omitted the results for gEEM because the algorithm is not publicly available. It is clear from the table that the proposed methods yield significantly lower mean angular differences (paired t-test, $p < 0.001$) with lower standard deviations than incremental gEEM. Similar to Table 1, gradient descent with MILP initialization gives the best result.

[3] `http://www.emmanuelcaruyer.com/q-space-sampling.php`

Table 2. Angular differences between estimated and ground-truth fiber directions for the sampling schemes generated by different methods

	Incr. CSC	Incr. CSC + Grad. Desc.	MILP	MILP + Grad. Desc.	Incr. gEEM
Angular Difference	$1.44^\circ \pm 0.69^\circ$	$1.40^\circ \pm 0.69^\circ$	$1.30^\circ \pm 0.68^\circ$	$\mathbf{1.29}^\circ \pm \mathbf{0.68}^\circ$	$1.72^\circ \pm 0.79^\circ$

4 Conclusion

To our knowledge, this is the first work on designing single-/multi-shell sampling schemes using continuous spherical code (CSC) and discrete spherical code (DSC) formulations. We propose an incremental estimation method for both CSC and DSC, a mixed-integer linear programming (MILP) method for DSC, and a Riemannian gradient descent method for CSC. The experimental results showed that, compared with the gEEM method and its incremental variant that has been used in the HCP, the sampling schemes by the proposed gradient descent with MILP initialization and incremental CSC yield larger covering radius and better rotation invariance.

Acknowledgement. This work was supported in part by a UNC start-up fund and NIH grants (EB006733, EB008374, EB009634, MH088520, AG041721, and MH100217).

References

1. Jones, D.K., Horsfield, M.A., Simmons, A.: Optimal strategies for measuring diffusion in anisotropic systems by magnetic resonance imaging. Magnetic Resonance in Medicine 42, 515–525 (1999)
2. Cook, P., Bai, Y., Nedjati-Gilani, S., Seunarine, K., Hall, M., Parker, G., Alexander, D.: Camino: Open-source diffusion-MRI reconstruction and processing. In: ISMRM (2006)
3. Caruyer, E., Cheng, J., Lenglet, C., Sapiro, G., Jiang, T., Deriche, R.: Optimal Design of Multiple Q-shells experiments for Diffusion MRI. In: CDMRI Workshop, MICCAI 2011 (2011)
4. Caruyer, E., Lenglet, C., Sapiro, G., Deriche, R.: Design of multishell sampling schemes with uniform coverage in diffusion MRI. Magnetic Resonance in Medicine (2013)
5. Sotiropoulos, S.N., Jbabdi, S., Xu, J., Andersson, J.L., Moeller, S., Auerbach, E.J., Glasser, M.F., Hernandez, M., Sapiro, G., Jenkinson, M., et al.: Advances in diffusion MRI acquisition and processing in the Human Connectome Project. NeuroImage 80, 125–143 (2013)
6. Conway, J.H., Hardin, R.H., Sloane, N.J.: Packing lines, planes, etc.: Packings in Grassmannian spaces. Experimental Mathematics 5(2), 139–159 (1996)
7. Deriche, R., Calder, J., Descoteaux, M.: Optimal real-time q-ball imaging using regularized kalman filtering with incremental orientation sets. Medical Image Analysis 13(4), 564–579 (2009)
8. Tóth, L.F.: On the densest packing of spherical caps. The American Mathematical Monthly 56(5), 330–331 (1949)
9. Gurobi Optimization, Inc.: Gurobi optimizer reference manual (2014)
10. Cheng, J., Ghosh, A., Jiang, T., Deriche, R.: A Riemannian Framework for Orientation Distribution Function Computing. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009, Part I. LNCS, vol. 5761, pp. 911–918. Springer, Heidelberg (2009)
11. Cheng, J., Ghosh, A., Jiang, T., Deriche, R.: Model-free and Analytical EAP Reconstruction via Spherical Polar Fourier Diffusion MRI. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part I. LNCS, vol. 6361, pp. 590–597. Springer, Heidelberg (2010)

A Prototype Representation to Approximate White Matter Bundles with Weighted Currents

Pietro Gori[1], Olivier Colliot[1], Linda Marrakchi-Kacem[1,2], Yulia Worbe[1], Fabrizio De Vico Fallani[1], Mario Chavez[1], Sophie Lecomte[1,2], Cyril Poupon[2], Andreas Hartmann[1], Nicholas Ayache[3], and Stanley Durrleman[1]

[1] Inria Paris-Rocquencourt, Sorbonne Universités, UPMC Univ Paris 06 UMR S1127, Inserm U1127, CNRS UMR 7225, ICM, F-75013, Paris, France
[2] Neurospin, CEA, Gif-Sur-Yvette, France
[3] Asclepios project-team, Inria Sophia Antipolis, Sophia Antipolis, France

Abstract. Quantitative and qualitative analysis of white matter fibers resulting from tractography algorithms is made difficult by their huge number. To this end, we propose an *approximation scheme* which gives as result a more concise but at the same time exhaustive representation of a fiber bundle. It is based on a novel computational model for fibers, called *weighted currents*, characterised by a metric that considers both the pathway and the anatomical locations of the endpoints of the fibers. Similarity has therefore a twofold connotation: geometrical and related to the connectivity. The core idea is to use this metric for approximating a fiber bundle with a set of weighted prototypes, chosen among the fibers, which represent ensembles of similar fibers. The weights are related to the number of fibers represented by the prototypes. The algorithm is divided into two steps. First, the main modes of the fiber bundle are detected using a *modularity based clustering* algorithm. Second, a *prototype fiber selection* process is carried on in each cluster separately. This permits to explain the main patterns of the fiber bundle in a fast and accurate way.

1 Introduction

Tractography from diffusion-weighted magnetic resonance imaging (DW-MRI) is a technique capable to virtually map the neural architecture of the human brain white matter (WM). This method is very useful for a better characterization of neurological diseases, for surgical planning or for the study of anatomo-functional relationships, for example. Tractography methods result in a set of 3D tracts that are commonly referred to as "fibers" which represent an estimate of the trajectory of large groups of neural fibers. WM fibers are traced starting from points, called seeds, inside one or more voxels and they are constituted by segments connecting different voxels. Tractography algorithms can be divided into two classes: deterministic and probabilistic. Deterministic methods reconstruct the fiber following the principal direction given by the diffusion model inside each voxel while the probabilistic ones use a randomly perturbed version of the main fiber direction. The idea behind probabilistic methods is to start many

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 289–296, 2014.

fibers from every seed in order to obtain maps of connectivity that are related to the probability that a certain voxel is connected to the starting seed. The choice of the starting voxels depends on the application and it ranges from the whole-brain to particular areas of the gray matter (ROI). One can also choose a second ROI as ending area in order to retrieve only the fibers connecting two precise areas of the gray matter. This paper focuses on such sets of fibers, called fiber bundles, which connect the cortical surface to the basal ganglia. A fiber is therefore characterised by its pathway between voxels and by the anatomical locations of its starting and ending point. Tractography algorithms usually result in a considerable amount of fibers. This makes difficult the development of efficient computational methods and the visualisation and interpretation of the brain connections that have been captured. To this end, we propose a *novel approximation scheme* for fiber bundles which reduces the number of tracts to analyse conserving almost all the information related to the fibers pathway (geometry) and to the distribution of the fibers endpoints (connectivity).

The core idea is to approximate a fiber bundle B with a set of weighted prototypes $\{\tau_i M_i\}$, chosen among the fibers, which represent ensembles of similar fibers. The weights $\{\tau_i\}$ are linked to the number of fibers approximated by the prototypes $\{M_i\}$. In order to do that, we need an appropriate dissimilarity measure based on both the geometry and the connectivity of a fiber. Usual dissimilarity measures present in the literature consider only the geometry of the fibers, like the Hausdorff distance [5], the Chamfer distance (or modified versions of it)[5,6,8,12], the total square loss between two fibers represented as Gaussian mixture model [13] or Fourier descriptors [10]. Other dissimilarity measures consider only the ending points of the fibers [11] or are based on the voxel space [12,14]. Here we propose a novel dissimilarity measure which considers both the pathway and the anatomical locations of the *termini* of the fibers. It can be seen as an extension of the framework of currents [3] and we have called it: *weighted currents*. As usual currents, it does not need point-to-point correspondence between fibers, it has a closed form and easy to compute expression and it can be used to compare fiber bundles. In addition, two fibers modelled as weighted currents are considered similar if and only if their pathways are alike and their endpoints are close to each other. Fibers are considered as vectors in the space of weighted currents. This permits to see a fiber bundle B as the sum of all its fibers F_j: $B = \sum_j F_j$ and its approximation as a weighted sum: $\sum_{i=1}^{K} \tau_i M_i$.

The proposed approximation scheme is based on the minimization of the residual error between B and $\{\tau_i M_i\}$ which can be easily written as: $||B - \sum_{i=1}^{K} \tau_i M_i||^2_{W_*}$ in the framework of weighted currents. The final goal is to find the smallest set of prototypes which minimizes this approximation error. An exhaustive analysis of all the possible combinations of prototypes is not feasible and therefore we propose a greedy algorithm. Our strategy is first to decompose a fiber bundle into modes, and then approximate every mode assuming that its fibers follow a gaussian distribution. In order to do that, the fiber bundle is first divided into independent clusters, which can be seen as different modes, using a *clustering algorithm*. Afterwards, the *prototype fiber selection* process is carried

on in each cluster separately. This permits a fast (parallel computing), concise and at the same time exhaustive approximation of the fiber bundle.

Usual clustering approaches for WM tracts are based on hierarchical [8,5,12,13], spectral [6,11] or EM [7] techniques. In many cases one needs to fix in advance parameters like the number or the size of the clusters. A solution, which does not require parameter setting, is to employ one of the quality functions in the field of networks community detection: the *modularity* [9]. It determines how good a cluster division is by comparing the similarity of fibers inside clusters with respect to the similarity of fibers between clusters. Optimization methods based on the maximization of this quality function permit to automatically find the number of clusters and how to divide the fiber bundle into clusters.

Eventually, the prototypes are chosen by selecting the fibers that minimize the residual error between B and $\{\tau_i M_i\}$ in the spirit of PCA (Principal Component Analysis) exploiting the fact that every mode is assumed to be gaussian (unimodal). This permits to explain and approximate almost the entire variability of the fiber bundle in a fast and accurate way.

2 Approximation Scheme for WM Fiber Bundles

In the first paragraph we introduce the concept of weighted currents from a formal mathematical point of view. Afterwards, we present the two steps of the approximation scheme based on weighted currents: the modularity optimization clustering algorithm and the prototype fiber selection.

Weighted Currents can be seen as an adaptation of the functional currents presented in [4] where the "functional signal" attached to each fiber is the anatomical location of its starting and ending point. Let X be a WM tract (an oriented and rectifiable curve in $\boldsymbol{R}^3$) which can be modelled as a polygonal line of N segments and f^c and f^b two 3D vectors in $M{=}\boldsymbol{R}^3\mathrm{x}\boldsymbol{R}^3$ containing the coordinates of the two extremities. For example, f^c might represent the extremity of the fiber on the cortical surface and f^b on the basal ganglia. In practice, f^c and f^b could be the mean values of the last n points, i.e. $n{=}3$, if the extremities are supposed to be in an area with a low SNR. The tract X, exactly as for currents, can be seen as a discrete weighted current via: $C_X(w) = \sum_{i=1}^{N} w(x_i, f^c, f^b)^T \alpha_i$ where x_i and α_i are the centres and the tangent vectors of the N segments respectively and w is a vector field belonging to a reproducing kernel Hilbert space (RKHS) W on $\boldsymbol{R}^3\mathrm{x}M$. A natural way to build the kernel associated to W is as the tensor product of kernels defined separately in the geometrical space $\boldsymbol{R}^3$ and in the "functional" space M. This means that the inner product in the framework of weighted currents between two tracts X ($C_X(w) = \sum_{i=1}^{N} w(x_i, f^c, f^b)^T \alpha_i$) and Y ($C_Y(w) = \sum_{j=1}^{M} w(y_j, t^c, t^b)^T \beta_j$) can be defined as:

$$\langle C_X, C_Y \rangle_{W*} = K_c(f^c, t^c) K_b(f^b, t^b) \underline{\sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_i^T K_g(x_i, y_j) \beta_j}$$

where K_c, K_b and K_g are Gaussian kernels parametrized by their scale parameters: λ_c, λ_b and λ_g respectively. The underlined part is the inner product

between usual currents which measures overall differences in the pathway of the two fibers. λ_g defines the range of interactions between the points of X and Y. The multiplication by the new terms K_c and K_b means that two fibers are considered similar if the pathways are similar, like in usual currents, but also if the endpoints of the two fibers are at a distance smaller than λ_c on the cortical surface and than λ_b on the basal ganglia (see Fig.1). The space of weighted currents is a vector space which implies that a fiber bundle B is seen as the sum of its fibers F_i: $C_B = \sum_i^N C_{F_i}$ and that it is also possible to compute the mean of a fiber bundle: $C_{\bar{F}} = \frac{1}{N}\sum_i^N C_{F_i}$. In the following, we will assume that each fiber is modelled as weighted current writing simply F instead than C_F. As we will see, our processing will be calculated using only simple computations of the Gram matrix $\Gamma = \{\langle F_i, F_j \rangle_{W*}\}_{i,j=1,\ldots,N}$.

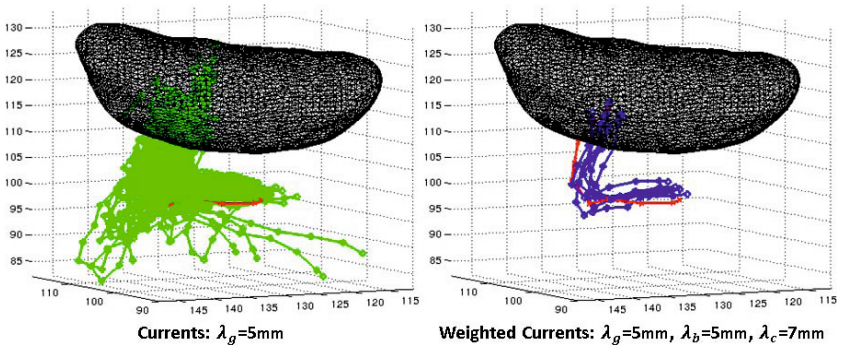


Fig. 1. Tracts that have an angle smaller than 45 degrees with the red one using currents (green, #118) and weighted currents (blue, #8). Fibers belong to a thalamo-cortico bundle of the right hemisphere resulting from a probabilistic tractography (see Sec.3). Green tracts are more than the blue ones and they are also more spread, connecting anatomical locations far from the ones of the red fiber. This shows why weighted currents are more suitable for clustering and approximating fiber bundles.

Modularity Optimization Clustering. A fiber bundle is seen as a sum of vectors in the space of weighted currents, each vector representing a fiber. Every vector can then be considered as a vertex of a weighted graph where the weighted edges are the inner products between every couple of fibers. The first step of the proposed algorithm consists in finding clusters in this set of vectors representing the main modes of the fibers distribution. We use a clustering algorithm based on the maximization of a quality function Q called Modularity [9]:

$$Q = \sum_{c=1}^{N_C} \left\{ \| \sum_{i \in c} F_i \|_{W^*}^2 \| \sum_{j \notin c} F_j \|_{W^*}^2 - (\sum_{i \in c} \sum_{j \notin c} \langle F_i, F_j \rangle_{W^*})^2 \right\} \qquad (1)$$

where N_C is the number of clusters. This equation can be easily rewritten in terms of the Gram matrix Γ as: $Q = \sum_{c=1}^{N_C} (s_c^T \Gamma s_c)((1\text{-}s_c)^T \Gamma (1\text{-}s_c)) - (s_c^T \Gamma (1 - s_c))^2$ where

$s_c(k)$=1 if F_k belongs to cluster c and 0 otherwise. In the simple case of $N_C = 2$ this equation can be rewritten as: $Q = \|\bar{F}_1\|^2_{W^*}\|\bar{F}_2\|^2_{W^*} - \langle \bar{F}_1, \bar{F}_2 \rangle^2_{W^*}$ where $\bar{F}_1$ and $\bar{F}_2$ are the means of the two clusters. Maximizing Q means therefore looking for two clusters whose means are as orthogonal as possible and at the same time their norms should be as close as possible. This can be generalized to N_C clusters by saying that the goal of modularity is to create clusters with balanced norms characterised by fibers orthogonal to the fibers in the other clusters and parallel to the fibers in their own cluster. Unfortunately, exact modularity optimization is a NP-complete problem. The "Louvain" algorithm [9] is a greedy solution divided into two steps which are repeated iteratively. At the beginning every WM tract forms a different cluster. The first part consists of associating every WM tract to all its neighbour clusters finding the one that leads to the largest increase in Q. This step is repeated until no change would produce an increase in Q. The second part consists in merging all the WM tracts of one cluster in one single supervertex. Two supervertices have a weighted edge equal to the sum of all the inner products between the fibers of the initial clusters. The two steps are repeated until no change would produce an increase in Q. At the end of this process the fiber bundle is separated into different clusters without fixing in advance neither the number of clusters nor their size.

Prototype Fiber Selection. The goal of the Prototype Fiber Selection (PFS) process is to concisely represent the fiber bundle B with a set of weighted prototypes $\{\tau_i M_i\}$ chosen among the fibers. If we wanted only one weighted prototype $\tau_1 M_1$ which minimizes $||B - \tau_1 M_1||^2_{W_*}$, it would be: $M_1 = \text{argmax}_{F_z} \langle B, \frac{F_z}{||F_z||} \rangle^2_{W_*}$ $= \text{argmax}_{F_z} N^2 \langle \bar{F}, \frac{F_z}{||F_z||} \rangle^2_{W_*}$ with $\tau_1 = \frac{\langle B, M_1 \rangle}{||M_1||^2}$. This means that we would look for the fiber most similar to the average of the bundle. This scheme works fine only in a uni-modal setting (i.e. gaussian) but not in a multi-modal one since the fiber chosen would be the one closest to the center of the different modes. If the modes are far from each other the fiber chosen could be also considered as an outlier. This is why it is fundamental finding the main modes of the bundle through the previous clustering step. Once defined the main modes, a PFS is performed independently on each one of them. One prototype is not sufficient to explain the whole cluster. So, as for instance in PCA, we remove from each fiber (F_i) its orthogonal projection onto the prototype $(\pi(F_i) = \frac{\langle F_i, M_1 \rangle M_1}{||M_1||^2})$ and we select, in this new representation $(r(F_i) = F_i - \pi(Fi))$, the fiber most similar to the new average as second prototype $(M_2 = \text{argmax}_{r(F_z)} N^2 \langle r(\bar{F}), \frac{r(F_z)}{||r(F_z)||} \rangle^2)$. We iterate this process for each cluster C_j until: $||C_j - \sum_{i=1}^{K_j} \tau_i M_i||_{W_*} \leq \gamma ||C_j||_{W_*}$. At iteration k, the k weights $\{\tau_i\}$ are computed by the orthogonal projection of the cluster C_j to the space spanned by the selected k prototypes $\{M_i\}$. It is important to notice that all these computations are based on the Gram matrix Γ of the fiber bundle, also when computing a new prototype: $\langle r(F_i), r(F_j) \rangle = \langle F_i, F_j \rangle - \frac{\langle F_i, M \rangle \langle F_j, M \rangle}{||M||^2} = \Gamma_{ij} - \frac{\Gamma_{iM} \Gamma_{jM}}{||\Gamma_{MM}||^2}$. After selecting the prototypes in each cluster, the weights are recomputed by the orthogonal projection of the whole bundle B to the entire set of prototypes in order to retrieve the correct values also for the prototypes close to the boundary between two different clusters.

3 Experiments and Discussion

We illustrate the algorithm on 36 fiber bundles from 6 subjects connecting the cortex to thalamus (12), putamen (12) and caudate (12) of the right hemisphere using both deterministic and probabilistic tractography estimated from HARDI data ([1] and references therein). The fiber bundles include the commisural fibers which have been truncated at the interhemispheric fissure [1]. The segmentation of the sub-cortical structures and of the cortex is done using FSL and FreeSurfer respectively [2]. The result of the clustering and the prototype representation for both deterministic and probabilistic tractography on a cortico-thalamus fiber bundle are shown in Fig.2. Other results are presented in the supplementary material. In the deterministic case only 491 fibers out of 9525 have been chosen as prototypes (∼5%) in order to explain 90% of $||B||_{W_*}$ while with the probabilistic tractography only 2652 fibers out of 10410 have been used as prototypes (∼25%) to explain 85% of $||B||_{W_*}$. The average reduction in the deterministic case is 4.7% explaining 90% of $||B||_{W_*}$ and 24.6% in the probabilistic case explaining 84% of $||B||_{W_*}$. Once calculated the Gram matrix (∼120 min), the computation of the whole approximation scheme is very fast: about 40 sec for the clustering and 30 sec for the PFS in a PC, Intel Xeon, 4 cores, 3.20GHz using 10410 fibers. Fig.3 shows the same bundle of Fig.2 but its goal is to point out the differences between our approximation scheme and a uniform downsampling resulting in the same number of tracts. The probability densities of the endpoints of the original fibers on both cortex and thalamus are more similar to the ones of our scheme than with a uniform downsampling. In fact, some small modes are lost with the uniform downsampling. Probability densities have been computed using gaussian kernels, taking into account the weights τ_i of the prototypes

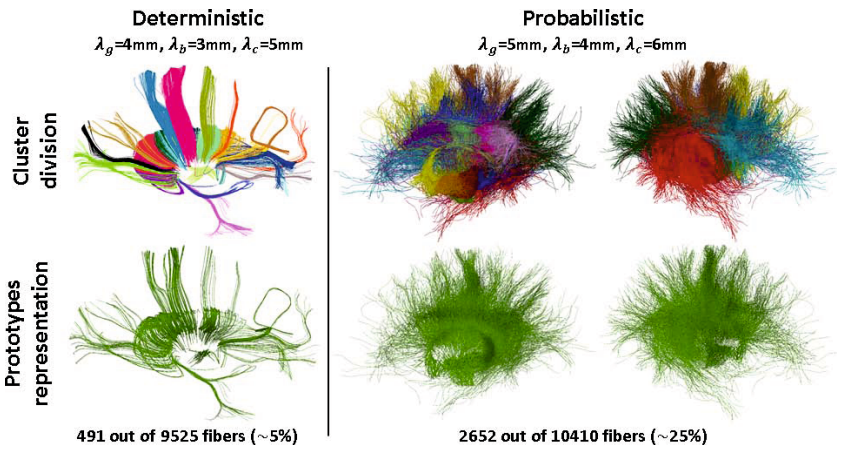


Fig. 2. First row: clusters of the right thalamo-cortico bundle highlighted in different colors. Second row: resulting prototypes visualized as tubes whose radii are proportional to their weights τ_i. The two columns on the right show the same fiber bundle from both lateral views.

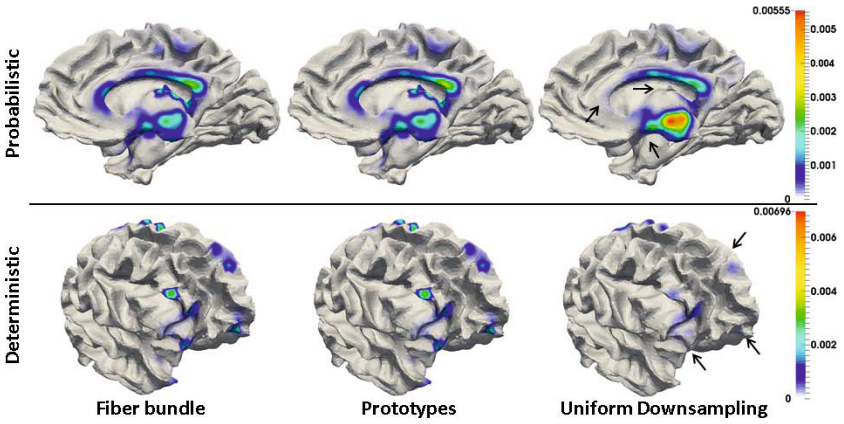


Fig. 3. Probability densities of the endpoints of the fibers using the right thalamo-cortico fiber bundle, prototypes P and an equal number of fibers coming from a uniform downsampling U of the original fiber bundle. U is a poorer approximation of the original density than P since it reveals fewer modes, as shown by the arrows.

for our approximation. The Kolmogorov-Smirnov test fails to show statistically significant differences between the densities of the original fibers and of the weighted prototypes and in most of the cases it finds significant differences at the 5% level between densities from original fibers and uniform downsampling.

4 Conclusions

We have presented here a new approximation scheme for white matter fiber bundles which results in a concise representation maintaining almost all the information related to the pathways of the fibers and to the locations of the fiber endpoints. It is based on a new computational model for fiber bundles, called weighted currents, which permits to compare fibers considering both their pathways and the locations of their endpoints. Moreover, it allows to treat fibers as vectors. We have tested this method on fiber bundles resulting from both deterministic and probabilistic tractography showing that the number of tracts to analyse can be reduced up to 3% of the initial number of fibers in the deterministic case and up to 17% in the probabilistic case, explaining in both cases more than 84% of the norm of the fiber bundles. We have also shown that the connectivity information, seen as the probability density of the fibers endpoints onto the gray matter, is more preserved with the proposed method than using a uniform downsampling. This information is important in many neuro-anatomical studies since it could be used, for instance, to characterise neurodevelopmental disorders [2]. Future works will include this new representation into registration and atlas construction methods in order to find changes in white matter organization across subjects. Another possible development could be to use this method for the white matter segmentation problem [6,12].

Acknowledgements. The research leading to these results has received funding from the program "Investissements d'avenir" ANR-10-IAIHU-06.

References

1. Marrakchi-Kacem, L., Delmaire, C., Guevara, P., Poupon, F., Lecomte, S., Tucholka, A., Roca, P., Yelnik, J., Durr, A., Mangin, J., Lehéricy, S., Poupon, C.: Mapping Cortico-Striatal Connectivity onto the Cortical Surface: A New Tractography-Based Approach to Study Huntington Disease. PLoS One 8, e53135 (2013)
2. Worbe, Y., Gerardin, E., Hartmann, A., Valabrègue, R., Chupin, M., Tremblay, L., Vidailhet, M., Colliot, O., Lehéricy, S.: Distinct structural changes underpin clinical phenotypes in patients with Gilles de la Tourette syndrome. Brain 133, 3649–3660 (2010)
3. Durrleman, S., Fillard, P., Pennec, X., Trouvé, A., Ayache, N.: Registration, Atlas Estimation and Variability Analysis of White Matter Fiber Bundles Modeled as Currents. NeuroImage 55, 1073–1090 (2011)
4. Charon, N., Trouvé, A.: Functional Currents: A New Mathematical Tool to Model and Analyse Functional Shapes. J. Math. Imaging Vis. 48, 413–431 (2014)
5. Gerig, G., Gouttard, S., Corouge, I.: Analysis of Brain White Matter via Fiber Tract Modeling. In: 26th IEEE EMBS, vol. 2, pp. 4421–4424. IEEE Press, New York (2004)
6. O'Donnel, L.J., Westin, C.F.: Automatic Tractography Segmentation Using a High-Dimensional White Matter Atlas. IEEE Trans. Med. Imaging 26, 1562–1575 (2007)
7. Maddah, M., Grimson, W., Warfield, S., Wells, W.: A unified framework for clustering and quantitative analysis of white matter fiber tracts. Med. Image Anal. 12, 191–202 (2008)
8. Zhang, S., Correia, S., Laidlaw, D.: Identifying White-Matter Fiber Bundles in DTI Data Using an Automated Proximity-Based Fiber Clustering Method. IEEE Trans. Vis. Comput. Graph. 14, 1044–1053 (2008)
9. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 10, P10008 (2008)
10. Batchelor, P.G., Calamante, F., Tournier, J., Atkinson, D., Hill, D.L., Connelly, A.: Quantification of the shape of fiber tracts. Magn. Reson. Med. 55, 894–903 (2006)
11. Brun, A., Park, H.J., Knutsson, H., Westin, C.F.: Coloring of DT-MRI Fiber Traces Using Laplacian Eigenmaps. In: Moreno-Díaz Jr., R., Pichler, F. (eds.) EUROCAST 2003. LNCS, vol. 2809, pp. 518–529. Springer, Heidelberg (2003)
12. Guevara, P., Poupon, C., Rivire, D., Cointepas, Y., Descoteaux, M., Thirion, B., Mangin, J.: Robust clustering of massive tractography datasets. NeuroImage 54, 1975–1993 (2011)
13. Liu, M., Vemuri, B.C., Deriche, R.: Unsupervised Automatic White Matter Fiber Clustering Using a Gaussian Mixture Model. In: 9th IEEE ISBI, pp. 522–525. IEEE Press, New York (2012)
14. Tunç, B., Smith, A.R., Wasserman, D., Pennec, X., Wells, W.M., Verma, R., Pohl, K.M.: Multinomial probabilistic fiber representation for connectivity driven clustering. In: Gee, J.C., Joshi, S., Pohl, K.M., Wells, W.M., Zöllei, L. (eds.) IPMI 2013. LNCS, vol. 7917, pp. 730–741. Springer, Heidelberg (2013)

Hole Detection in Metabolic Connectivity of Alzheimer's Disease Using $k-$Laplacian

Hyekyoung Lee[1], Moo K. Chung[2], Hyejin Kang[1], and Dong Soo Lee[1]

[1] Dept. of Nuclear Medicine, Seoul National University, Republic of Korea
[2] Depart. of Biostatistics and Medical Informatics, University of Wisconsin, Madison, USA
hklee.brain@gmail.com, mkchung@wisc.edu, {hkang211,dsl}@snu.ac.kr

Abstract. Recent studies have found that the modular structure of functional brain network is disrupted during the progress of Alzheimer's disease. The modular structure of network is the most basic topological invariant in determining the shape of network in the view of algebraic topology. In this study, we propose a new method to find another higher order topological invariant, hole, based on persistent homology. If a hole exists in the network, the information can be inefficiently delivered between regions. If we can localize the hole in the network, we can infer the reason of network inefficiency. We propose to detect the persistent hole using the spectrum of $k-$Laplacian, which is the generalized version of graph Laplacian. The method is applied to the metabolic network based on FDG-PET data of Alzheimer disease (AD), mild cognitive impairment (MCI) and normal control (NC) groups. The experiments show that the persistence of hole can be used as a biological marker of disease progression to AD. The localized hole may help understand the brain network abnormality in AD, revealing that the limbic-temporo-parietal association regions disturb direct connections between other regions.

1 Introduction

The hierarchical modular structure of brain network has revealed the functional integration of local specialized modules of brain regions [1]. The modular structure of network is the first basic topological invariant in determining the shape of network in the view of algebraic topology [2]. The second basic topological invariant is holes. While the connected network structures of brain has been often studied, holes never played any role in modeling brain networks [3,1]. However, hole detection has found its usefulness in mobile sensor networks in determining the obstacle-regions, which weaken the strength of cellphone signals [4,5]. In this study, we take a novel hole detection method in finding such aberrant regions of brain network in Alzheimer's disease (AD).

If the brain network has the hole, it implies that the information can be inefficiently transferred between regions due to indirect connections around the hole. When abnormal brain regions associated with Alzheimer's disease interrupt direct connections between other regions, the hole can occur in the network. The

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 297–304, 2014.

larger the hole is, the more inefficient the information transfer around the hole is. Hence, the size of hole can be a new measure for quantifying the degree of abnormality of the brain network in AD [6]. To find holes and estimate their size, we introduce the concept of persistent homology which assumes that true topological invariants of the underlying network are more persistent over the change of network parameters rather than noise. The more persistent hole is over the change of network parameters, the more connections are needed to cover the hole. Thus, the persistence of hole can be considered as its size. The hole is usually identified by manipulating a matrix associated with the boundary operator in the persistent homology [7]. This method directly selects the edge set that forms a hole. However, this approach has an ambiguity in choosing edges that depends on the order of nodes and edges. A superior new method, which this paper is proposing, is to estimate holes by computing the eigenvectors with zero eigenvalues of higher order Laplacians [5], called k-Laplacian. The method represents the hole as a linear combination of edges of which coefficients are proportional to their contributions to the hole.

The methodological contributions of this paper are: (1) We propose a new method in detecting the local abnormality of network by identifying a hole within the persistent homology framework. This is the first study of using the hole as a brain network feature. (2) We introduce the concept of $k-$Laplacian in estimating the hole. This approach is a natural generalization of finding modular structure of brain network using the spectrum of graph Laplacian, i.e., $0-$Laplacian. (3) We demonstrate that the persistence of holes in the network can be used to quantify the disease progression for the first time. The proposed hole detection method is applied to the FDG-PET based metabolic network of AD, mild cognitive impairment (MCI) and normal control (NC) groups. Our finding suggests that the persistence of hole may be increased as the disease progressed. The resulting holes support prior studies that reported alterations and disconnections in temporal, parietal, frontal association areas and abnormal change in the limbic region by AD [8,9]. In addition, the result also show that the medial temporal lobe is affected by MCI [9].

2 Methods

2.1 $k-$Dimensional Holes

We will first define a hole in a network rigorously using the language of algebraic topology. The nodes and edges of a network are the building blocks of topological space defined on the network. The algebraic topology extends this concept further to a simplicial complex, which considers higher order elements with more than three nodes such as triangles. Given a node set $v_i \in V$, an element with $(k+1)$ nodes is called $k-$simplex $\sigma_k = [v_1, \ldots, v_{k+1}]$. Node, edge and filled-in triangle are then denoted as σ_0, σ_1 and σ_2. Note that we call σ_2 as the filled-in triangle in the simplicial complex to distinguish the unfilled-in triangle consisting of three nodes (σ_0) in a network. The collection of simplexes is a simplicial complex. The example of simplicial complex is shown in Fig. 1 (a).

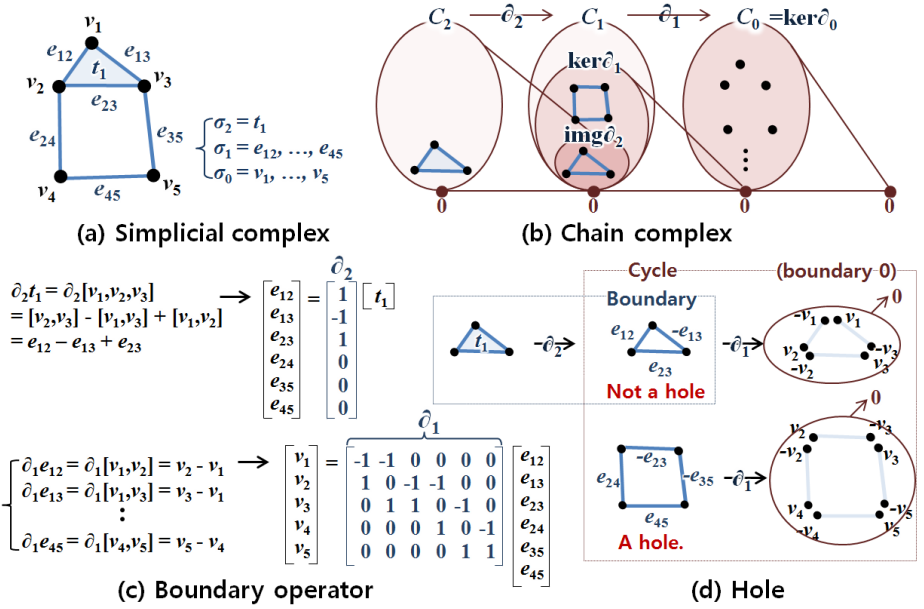


Fig. 1. (a) An example of simplicial complex. (b) A chain complex with chain, cycle and boundary groups which are mapped by boundary operator. (c) The boundary operators ∂_1 and ∂_2 of (a) in the matrix form. (d) A hole is the cycle whose boundary becomes zero, but not the boundary of any higher order simplex. Hence, the boundary of the filled-in triangle t_1 is not a hole but a cycle.

The boundary of an edge (σ_1) is two end nodes (σ_0) of the edge. The boundary of a filled-in triangle (σ_2) is three edges (σ_1) surrounding the triangle. If we denote C_k as a chain complex, a set of σ_ks, the relationship between σ_k and σ_{k-1} is defined using the boundary operator $\partial_k : C_k \to C_{k-1}$. Fig. 1 (b) shows a chain map by boundary operation. Given $\{\sigma_k^1, \cdots, \sigma_k^q\} \subset C_k$ and $\{\sigma_{k-1}^1, \cdots, \sigma_{k-1}^p\} \subset C_{k-1}$, the linear transformation $\partial_k \in \mathbb{R}^{p \times q}$ from C_k to C_{k-1} can be represented in the matrix form:

$$[\partial_k]_{ij} = \begin{cases} 1 \text{ if } \sigma_{k-1}^i \text{ is positively oriented w.r.t. } \sigma_k^j, \\ -1 \text{ if } \sigma_{k-1}^i \text{ is negatively oriented w.r.t. } \sigma_k^j, \\ 0 \text{ otherwise.} \end{cases}$$

When σ_{k-1}^i belongs to the ordered boundaries of σ_k^j, it is positively/negatively oriented if its order is odd/even. The matrix form of boundary operators is shown in Fig. 1 (c). ∂_1 is an incidence matrix of binary network. So the boundary operator is the generalization of the incident matrix in graph theory.

The kernel of ∂_k is a set defined by $\ker\partial_k = \{\sigma_k \in C_k | \partial_k \sigma_k = 0\}$. The kernel is then a cycle which consists of σ_ks starting and ending at the same σ_{k-1}. The image of ∂_{k+1} is a set defined by $\text{img}\partial_{k+1} = \{\sigma_k \in C_k | \partial_{k+1}\sigma_{k+1} = \sigma_k, \sigma_{k+1} \in C_{k+1}\}$. Hence, the image of ∂_{k+1} is a boundary of σ_{k+1}, which is always a cycle, i.e. $\text{img}\partial_{k+1} \subset \ker\partial_k$. But the cycle may not be a boundary of σ_{k+1} as shown in Fig. 1 (d), where the square hole is not the boundary of any higher order simplex. The set

of cycles of ∂_k, which are not the boundary of ∂_{k+1} is called the kth homology group. $H_k = \ker\partial_k \cap (\mathrm{img}\partial_{k+1})^C$, where $(\cdot)^C$ denotes the complementary set of $\cdot$ [7]. The element of H_k is the $k-$dimensional hole which is an important topological invariant used in distinguishing different topological spaces. The cardinality of H_k is the kth Betti number β_k. For the sake of simplicity, we will only consider 1-dimensional hole as a hole in this study and left higher dimensional holes as a future study.

2.2 $k-$Laplacian

In the persistent homology, the $k-$dimensional hole H_k is usually identified by manipulating the kernel of ∂_k and image of ∂_{k+1} based on Gaussian elimination [7]. If we apply this approach to the example in Fig. 1 (a), one of two possible holes, $e_{12}-e_{13}+e_{24}-e_{35}+e_{45}$ or $-e_{23}+e_{24}-e_{35}+e_{45}$, are estimated depending on how to order edges in the column of ∂_1 and row of ∂_2. To avoid this ambiguity, we introduce a new method based on $k-$Laplacian for estimating hole [10].

The k-Laplacian $\boldsymbol{L}_k$ is defined as $\boldsymbol{L}_k = \partial_{k+1}\partial_{k+1}^\top + \partial_k^\top\partial_k$ [11]. Since $\partial_0 : C_0 \to 0$, $\boldsymbol{L}_0 = \partial_1^\top\partial_1$ and it is the graph Laplacian, which is widely used in spectral clustering [12]. The kth homology group H_k is a kernel of $k-$Laplacian $\boldsymbol{L}_k$ [10]. Hence the kth Betti number β_k is the dimension of kernel space of $\boldsymbol{L}_k$. The eigenvectors with zero eigenvalues of $\boldsymbol{L}_k$ are spanned in the kernel space of $\boldsymbol{L}_k$. So, the $k-$dimensional hole in H_k is obtained by the eigenvectors with zero eigenvalues of $\boldsymbol{L}_k$. β_k is obtained by the number of zero eigenvalues of $\boldsymbol{L}_k$.

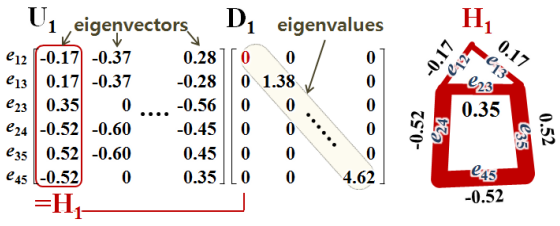


Fig. 2. Example of hole estimation based on eigenvector $\boldsymbol{U}_1$ and eigenvalue $\boldsymbol{D}_1$ of $\boldsymbol{L}_1$. The absolute value of eigenvector with zero eigenvalue becomes the edge weight in the hole H_1.

Fig. 2 shows the example of hole estimation using the spectrum of $\boldsymbol{L}_1$. The simplicial complex in Fig. 1 (a) is used as an example. After estimating $\boldsymbol{L}_1$ using ∂_1 and ∂_2 in Fig. 1 (c), we obtain its eigenvectors $\boldsymbol{U}_1$ and their corresponding eigenvalues $\boldsymbol{D}_1$. The eigenvector with zero eigenvalue of $\boldsymbol{L}_1$ is the hole H_1. The resulting hole can be represented in the linear combination of edges $-0.17e_{12} + 0.17e_{13} + 0.35e_{23} - 0.52e_{24} + 0.52e_{35} - 0.52e_{45}$ as shown in Fig. 2. The absolute value of coefficient is proportional to its contribution to the hole.

2.3 Persistent Holes

The metabolic brain connectivity forms the connectivity matrix $\boldsymbol{W} = [w_{ij}]$, with each of the elements w_{ij} encoding the distance between two brain regions v_i and v_j. We introduce the Rips complex to estimate holes in brain network with the connectivity matrix $\boldsymbol{W}$. The Rips complex $\mathcal{R}(\boldsymbol{W}, \epsilon)$ is a simplicial complex whose k-simplexes correspond to unordered $(k+1)$-tuples of nodes which are pairwise within distance ϵ [3]. If we confine $k \leq 1$, the Rips complex is identical

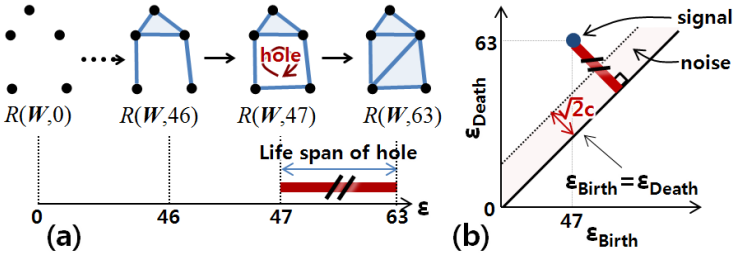


Fig. 3. (a) Rips filtration and (b) the persistence diagram of hole. Since the hole starts and ends at 47 and 63, respectively, the point is plotted at (47,63) on (b).

to the binary network where two nodes are connected if their distance is less than ϵ. Given the connectivity matrix $\boldsymbol{W}$ and thresholds $\epsilon_1 < \cdots < \epsilon_q$, the Rips filtration decomposes the connectivity matrix into the sequence of Rips complexes: $\mathcal{R}(\boldsymbol{W}, \epsilon_1) \subseteq \mathcal{R}(\boldsymbol{W}, \epsilon_2) \subseteq \cdots \subseteq \mathcal{R}(\boldsymbol{W}, \epsilon_q)$. During the filtration, the holes of Rips complexes are appearing and disappearing as shown in Fig. 3 (a). The persistent homology observes such a change of $k-$dimensional holes and counts their Betti numbers over the change of threshold. The birth and death times of hole ϵ_{Birth} and ϵ_{Death} are encoded in the persistence diagram by mapping to the point $(\epsilon_{Birth}, \epsilon_{Death})$. The persistence diagram $\mathcal{P}$ is a set of the points in the plane where the horizontal and vertical axes represent the birth and death times of hole as shown in Fig. 3 (b). The life span of hole from birth to death time is same as the distance from the point to the line $\epsilon_{Birth} = \epsilon_{Death}$ in the persistence diagram. The closer to the line $\epsilon_{Birth} = \epsilon_{Death}$, the shorter the life span of corresponding hole is. The persistent homology assumes that a persistent hole with long life span may be the signal that reflects the shape of true topological space, but a hole with short life span may be a noise. Now, we introduce the bottleneck distance to estimate the confidence band $[0, c]$ which distinguishes between signal and noise in Fig. 3 (b) [13].

The bottleneck distance W_∞ measures the distance between two persistence diagrams $\mathcal{P}_1$ and $\mathcal{P}_2$. It is defined as

$$W_\infty(\mathcal{P}_1, \mathcal{P}_2) = \min_P \max_{(x,y)\in P} d_\infty(x, y) \text{ for all } x \in \mathcal{P}_1, y \in \mathcal{P}_2,$$

where d_∞ is the L_∞ distance and P is a one-to-one correspondence of the points in $\mathcal{P}_1$ and $\mathcal{P}_2$. Suppose that the persistence diagrams of the underlying Rips and random Rips complexes $\mathcal{P}$ and $\mathcal{P}_{rand}$ are given. If we find a confidence interval c such that $P(W_\infty(\mathcal{P}, \mathcal{P}_{rand}) > c) \leq \alpha$, $\sqrt{2}c$ is same as the distance to the line $\epsilon_{Birth} = \epsilon_{Death}$ which distinguishes between the signal and noise [13]. The random Rips complexes are generated from 5000 random permutations of AD, MCI and NC datasets. For each permutation, the group labels are randomly reassigned and the persistence diagram $\mathcal{P}_{rand}$ and the bottleneck distance $T = W_\infty(\mathcal{P}, \mathcal{P}_{rand})$ are recalculated. After 5000 permutations, we obtain 5000 bottleneck distances and sort them in descending order as $T_1 > T_2 > \cdots > T_{5000}$. $c = T_{250}$ is chosen to satisfy $\alpha = 0.05$. Then, the hole located outside of the confidence band $[0, c]$ is denoted as the persistent hole. Here we expect that finding persistent holes may help understand aberrant functional connectivity in AD.

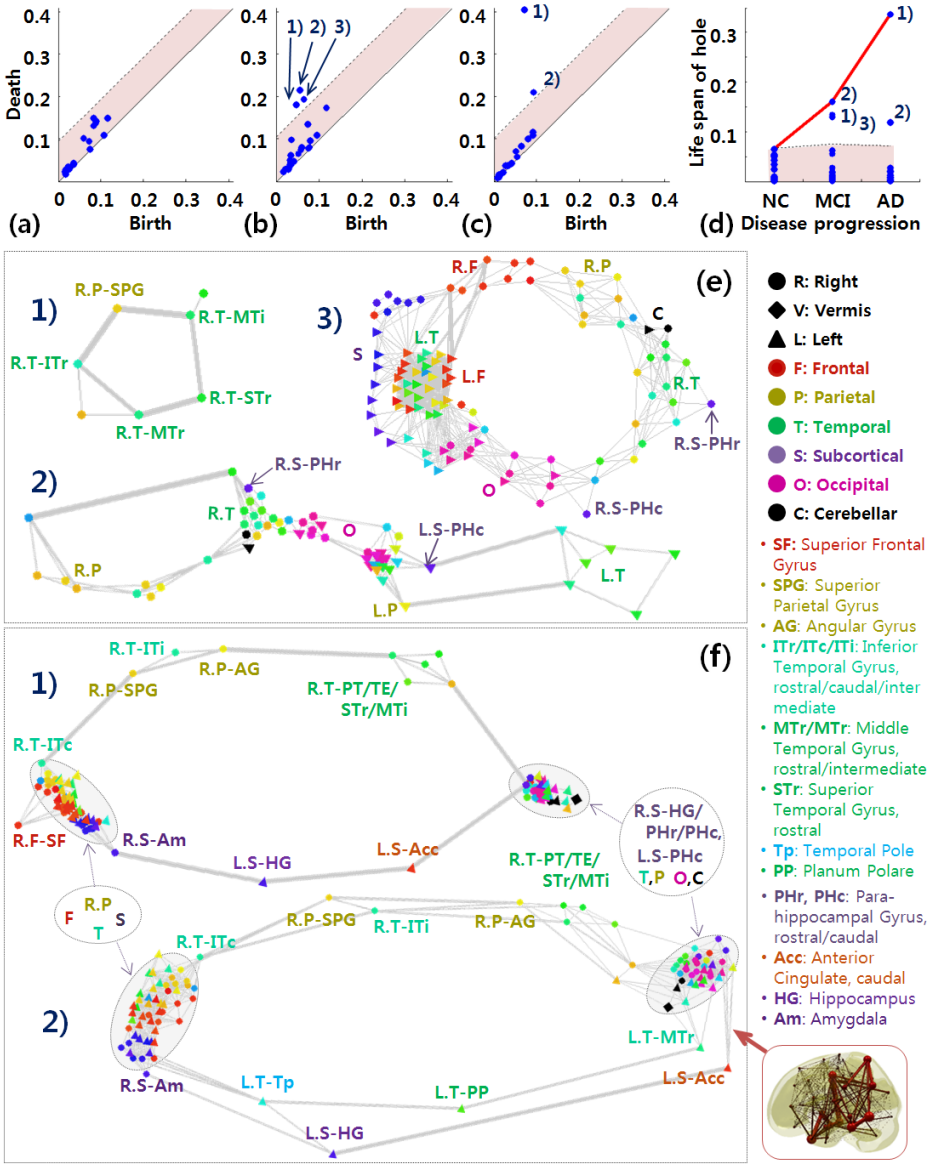


Fig. 4. The persistence diagram of (a) NC, (b) MCI and (c) AD. (d) The life span of hole with respect to the disease progression. There is no significantly persistent hole in NC. (e) Three persistent holes of MCI are mainly located in 1) right temporal, 2) bilateral temporo-parietal, and 3) widespread fronto-temporo-parietal-occipital lobule. (f) Two persistent holes of AD may occur because the limbic-temporo-parietal association regions disturb direct connections between two large clustered regions.

3 Results

Data Sets. We used FDG-PET imaging data sets: 45 NC (age: 68.9 ± 5.2), 24 MCI (67.8 ± 9.0) and 22 AD (66.9 ± 7.1) subjects. All ^{18}F-FDG PET images were spatially normalized and smoothed with 16 mm FWHM using the SPM package. Then, FDG uptake values of 103 regions of interest (ROIs) were extracted by weighted averaging. Each FDG uptake value was scaled by individual's total gray matter mean count. The connectivity matrix $\boldsymbol{W} = [w_{ij}] \in \mathbb{R}^{103 \times 103}$ was estimated based on the diffusion distance on positive correlation between FDG uptake values in two ROIs.

Group Differences. The persistence diagrams of NC, MCI and AD are shown in Fig. 4 (a-c). The total number of holes is 17 for AD, 25 for MCI and 21 for NC. We examined group differences using the bottleneck distance between persistence diagrams and permutation test. The persistence diagrams are significantly different between NC and AD ($p < 0.05$), but tend to be different between NC and MCI and between MCI and AD ($p < 0.1$).

Life Span of Holes. In Fig. 4 (d), the life span of holes is plotted with respect to NC, MCI and AD. Using the permutation test, we found that the longest life span of holes connected by red line is proportional to the disease progression ($p < 0.05$). The resulting confidence band for persistence diagram is shown in the shaded region in Fig. 4 (a-d). All holes in NC are not persistent. 2 and 3 holes are determined as the persistent for AD and MCI respectively.

Persistent Hole. 5 persistent holes are shown in Fig. 4 (e,f). Three persistent holes of MCI are mainly located in right temporal, bilateral temporo-parietal, and widespread fronto-temporo-parietal-occipital lobule. The reduced metabolism in a network has been found in parietal, temporal and frontal lobes in AD [8]. Especially, the hypometabolism of medial temporal lobe observed in the first hole is known as a biomarker for the identification of MCI [9]. In the persistent holes of AD, we found that two large clustered brain regions on the left and right sides in (f) are not directly connected because the limbic-temporo-parietal association regions disturb the connection between them. These association regions are also known to be affected by AD [8,9].

4 Conclusions

In this study, we propose a new method for localizing aberrant regions by detecting hole in the metabolic network based on persistent homology and 1-Laplacian. We also introduce a new biomarker, life span of hole, to measure the degree of abnormality of brain network. The proposed hole detection method is natural extension of finding modular structure of network based on the spectrum of graph Laplacian, $0-$Laplacian. The resulting aberrant holes are mainly located in parietal, temporal and frontal regions which is known to be related to AD and MCI. In addition, our finding suggests that the brain network inefficiency in AD

may be because the limbic-temporo-parietal association regions interrupt direct connections between other brain regions. The proposed method can be further applicable to other high order topologically invariant features using k-Laplacian, which is left as a future study.

Acknowledgments. This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2064593), by NRF grant funded by the Korea government (MEST) (2011-0030815), by a grant of the future-based technology development programof the NRF funded by the MEST (20100028755) and by NIH grant UL1TR000427 and Vilas Associate Award from Univ. of Wisconsin.

References

1. Park, H., Friston, K.: Structural and functional brain networks: from connections to cognition. Science 342(6158), 1238411 (2013)
2. Carlsson, G.: Topology and data. B. Am. Math. Soc. 46, 255–308 (2009)
3. Lee, H., Chung, M.K., Kang, H., Kim, B.N., Lee, D.S.: Persistent brain network homology from the perspective of dendrogram. IEEE T. Med. Imaging 31, 2267–2277 (2012)
4. de Silva, V., Ghrist, R.: Coverage in sensor networks via persistent homology. Algebraic and Geometric Topology 7, 339–358 (2007)
5. Muhammad, A., Egerstedt, M.: Control using higher order laplacians in network topologies. In: Proceedings of the 17th International Symposium on Mathematical Theory of Networks and Systems, pp. 1024–1038 (2006)
6. Daianu, M., Jahanshad, N., Nir, T.M., Toga, A.W., Jack, C.R., Weiner, M.W., Thompson, P.M.: Breakdown of brain connectivity between normal aging and alzheimer's disease: A structural k-core network analysis. Brain Connectivity 3(4), 407–422 (2013)
7. Edelsbrunner, H., Harer, J.: Computational Topology: An Introduction. American Mathematical Society Press (2009)
8. Alexander, G.E., Chen, K., Pietrini, P., Rapoport, S.I., Reiman, E.M.: PET evaluation of cerebral metabolic decline in dementia: A potential outcome measure in alzheimer's disease treatment studies. Am. J. Psychiatry 159, 738–745 (2002)
9. Mosconi, L., Tsui, W.H., De Santi, S., Li, J., Rusinek, H., Convit, A., Li, Y., Boppana, M., de Leon, M.J.: Reduced hippocampal metabolism in MCI and AD: automated FDG-PET image analysis. Neurology 64, 1860–1867 (2005)
10. Horak, D., Jost, J.: Spectra of combinatorial laplace operators on simplicial complexes. Advances in Mathematics 244, 303–336 (2013)
11. Awasthi, V.V.: A note on the application of incidence matrices of simplicial complexes. Int. Journal of Contemp. Math. Sciences 8(19), 935–939 (2013)
12. Chung, F.R.K.: Spectral Graph Theory. CBMS Regional Conference Series in Mathematics, No. 92. American Mathematical Society (1996)
13. Balakrishnan, S., Fasy, B.T., Lecci, F., Rinaldo, A., Singh, A., Wasserman, L.: Statistical inference for persistent homology (2013)

Deep Learning Based Imaging Data Completion for Improved Brain Disease Diagnosis

Rongjian Li[1], Wenlu Zhang[1], Heung-Il Suk[2], Li Wang[2], Jiang Li[3], Dinggang Shen[2], and Shuiwang Ji[1]

[1] Department of Computer Science, Old Dominion University, Norfolk, VA, USA
[2] Department of Radiology, University of North Carolina at Chapel Hill, NC, USA
[3] Department of ECE, Old Dominion University, Norfolk, VA, USA

Abstract. Combining multi-modality brain data for disease diagnosis commonly leads to improved performance. A challenge in using multi-modality data is that the data are commonly incomplete; namely, some modality might be missing for some subjects. In this work, we proposed a deep learning based framework for estimating multi-modality imaging data. Our method takes the form of convolutional neural networks, where the input and output are two volumetric modalities. The network contains a large number of trainable parameters that capture the relationship between input and output modalities. When trained on subjects with all modalities, the network can estimate the output modality given the input modality. We evaluated our method on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, where the input and output modalities are MRI and PET images, respectively. Results showed that our method significantly outperformed prior methods.

1 Introduction

Alzheimer's disease (AD) is a common neuro-degenerative disease for which we still lack effective treatment. It has been shown that early detection and intervention at its prodromal stage, such as the mild cognitive impairment (MCI) stage, are effective in delaying the onset of AD. Developments in neuroimaging techniques, such as the magnetic resonance imaging (MRI) and positron emission tomography (PET) techniques, coupled with advanced computational methods, have led to accurate prediction of AD and MCI [1].

A key challenge in employing computational methods for disease diagnosis is that the neuroimaging data usually consist of multiple modalities, but they could be incomplete in the sense that not all subjects have all data modalities. The accuracy of disease diagnosis might be improved if the missing data could be estimated. However, the relationship between different data modalities is complicated and nonlinear. Thus, a highly sophisticated model is required for the collaborative completion of neuroimaging data.

Deep convolutional neural networks (CNNs) are a type of multi-layer, fully trainable models that are capable of capturing highly nonlinear mappings between inputs and outputs [2]. These models were originally motivated from

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 305–312, 2014.

computer vision problems and thus are intrinsically suitable for image-related applications. Deep CNNs have been successfully applied to a variety of applications, including image classification [2,3], segmentation [4], and denoising [5].

In this work, we propose to use deep CNNs for completing and integrating multi-modality neuroimaging data. Specifically, we designed a 3-dimensional (3-D) CNN architecture that takes one volumetric data modality as input and another volumetric data modality as its output. When trained end-to-end on subjects with both data modalities, the network captures the nonlinear relationship between two data modalities. This allows us to predict and estimate the output data modality given the input modality.

We applied our 3-D CNN model to predict the missing PET patterns from the MRI data. We trained our model on subjects with both PET and MRI data, where the MRI data were used as input and the PET data were used as output. The trained network contains a large number of parameters that encode the nonlinear relationship between MRI and PET data. We used the trained network to estimate the PET patterns for subjects with only MRI data. Results showed that our method outperformed prior methods on disease diagnosis.

2 Material and Methods

2.1 Data Preprocessing

The data used in this work were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. For each subject, the T1-weighted MRI was processed by correcting the intensity inhomogeneity followed by skull-stripping and cerebellum removing. In addition, each MRI was segmented into gray matter, white matter and cerebrospinal fluid and was further spatially normalized into a template space. In this work, the gray matter tissue density maps were used. The PET images were also obtained from ADNI, and they were rigidly aligned to the respective MR images. The gray matter tissue density maps and the PET images were further smoothed using a Gaussian kernel (with unit standard deviation) to improve the signal-to-noise ratio. To reduce the computational cost, we downsampled both the gray matter tissue density maps and PET images to $64 \times 64 \times 64$ voxels.

We used data for 830 subjects in the ADNI baseline data set. This data set was acquired from 198 AD patients, 403 MCI patients, which include 167 pMCI patients (who will progress to AD in 18 months) and 236 sMCI patients (whose symptom were stable and will not progress to AD in 18 months), and 229 healthy normal controls (NC). Out of these 830 subjects, more than half of them (432) do not have PET images. Thus, accurate completion of PET images for these subjects would improve the accuracy of disease diagnosis.

2.2 3-D Convolutional Neural Networks

Convolutional neural networks (CNNs) are a type of deep models that are able to capture highly nonlinear relationships between input and output [2]. In image

classification tasks, two types of layers, *i.e.*, convolutional layer and subsampling layer, are usually stacked alternatingly. The convolutional layer applies trainable filers to feature maps in the previous layer, while the subsampling layer is used to reduce the resolution of feature maps.

CNNs have been primarily applied to 2-D images such as visual object recognition [2,3] and segmentation [6]. In [4,5], 2-D CNNs have been extended to segment and restore 3-D images. In [7], 3-D CNNs have been applied to process spatiotemporal video data. Similar to the 2-D case, 3-D CNNs perform nonlinear mapping by computing convolutions with 3-D filters.

Formally, let the value at position (x, y, z) on the jth feature map in the ith layer be v_{ij}^{xyz}. Then the 3-D convolution is given by

$$v_{ij}^{xyz} = \sigma\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right), \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function, b_{ij} is the bias, m indexes the set of feature maps in the $(i-1)$th layer connected to the current feature map, P_i, Q_i and R_i are the sizes of the 3-D kernel along three spatial dimensions respectively, w_{ijm}^{pqr} is the (p, q, r)th value of the filter connected to the mth feature map in the previous layer. Note that Eq. (1) describes a generic 3-D convolution operation and can be applied to any layer of a 3-D CNN architecture with any number of feature maps.

Subsampling layers are commonly used in recognition and classification tasks. In these layers, the resolution of feature maps is reduced by pooling over local neighborhood, thereby enhancing invariance to distortions on the inputs. In this work, our primary focus is data completion instead of recognition. Thus, subsampling layers were not used.

2.3 3-D CNN for Imaging Data Completion

Based on the 3-D convolution described above, a variety of CNN architectures can be devised. In the following, we describe a 3-D CNN architecture, shown in Fig. 1, for PET image completion. The data for training this CNN model consist of patches extracted from subjects having both PET and MRI images. The input patch size was determined by the size of output patch in the output layer, since each convolution operation reduces the size of feature map along each dimension by a factor related to the size of filter. In this work, the size of output patches was set to $3 \times 3 \times 3$. We randomly selected a large number of patches from each 3-D MRI volume, and the corresponding PET image patches were also obtained. Patches that cross the boundary or are located completely within background were removed. The total number of patches extracted from each volume was $50,000$ so that the entire volume is largely covered.

In the CNN architecture, we first applied 3-D convolution with a filter size of $7 \times 7 \times 7$ on the input patch and construct 10 feature maps in the first hidden layer. The second hidden layer is again a 3-D convolution layer with 10

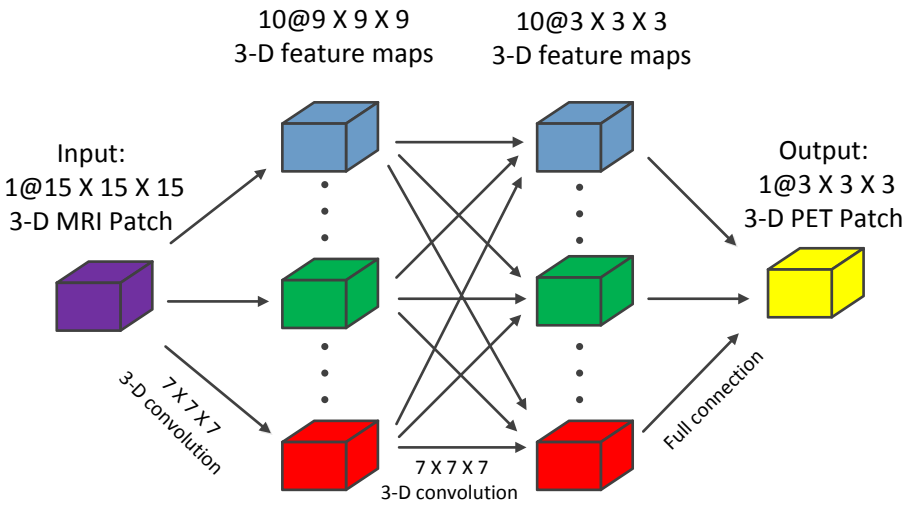


Fig. 1. The 3-D CNN architecture for imaging data completion used in this work. There are 2 hidden layers between the input and output layers. Each of the hidden layers contains 10 feature maps. The total number of trainable parameters in this network is 37, 761.

feature maps fully connected to all the feature maps in the previous layer. The output layer contains only one feature map, which is the corresponding PET image patch. In addition, the filter size for mapping the feature maps of the last hidden layer to the output was set to 1 to reduce the computational cost. In total, the number of trainable parameters for this network is 37, 761. The latent nonlinear relationship between the MRI and PET images was encoded into the large number of parameters in the network. This CNN architecture was selected based on a balance between the representation power and the computational cost of training the network. A network with more layers and feature maps might be able to represent the training data better, but the computational cost of training more complex networks is prohibitive.

In this work, the CNS package [8] was used to implement the CNN architecture. The weights of this network were updated by error back-propagation using stochastic gradient descent algorithm. The learning rate was fixed to 10^{-2} in all the experiments, and other parameters were set to the default values given in the CNS package [8]. The network was trained for multiple epochs, where each epoch involves training the network by each example once. In this paper, we trained the network for 10 epochs since the performance seems to have converged after 10 epochs and the training was very time-consuming. In particular, we have $398 \times 50,000 = 19.9$ million training patches. Each epoch took about 48 hours if all the patches were used on a Tesla K20c GPU with 2,496 cores.

3 Results and Discussion

3.1 Experimental Setup

In the experiments, we focused on evaluating our 3-D CNN model for missing PET data completion. We used several controlled experiments to compare the predicted and the true PET image data. We did not employ advanced feature extraction and classification methods to compare the completed and true data, but rather used a set of standard methods to make the comparison straightforward. We consider three binary-class classification tasks (*i.e.*, AD vs. NC, MCI vs. NC, and sMCI vs. pMCI) in this paper, where MCI includes both pMCI and sMCI.

We compared our method with two other commonly used missing data estimation methods, namely, K-nearest neighbor (KNN) and Zero methods [9]. The experiments in this work consist of two steps. The first step is to complete the missing PET data using CNN, KNN, or Zero methods. The second step then evaluate the classification performance based on reconstructed data using the ℓ_2-norm regularized logistic regression classifiers for all methods. In the experiments, we trained the classifiers by randomly selecting 2/3 of the samples and performed an evaluation using the remaining 1/3 as test data in the second step. To obtain robust performance estimates, we repeated the random partition 30 times and reported the statistics computed over these 30 trials. Note that no class information was used in our CNN training. Thus, we built one CNN model and applied it for all 30 random trials. We performed feature selection by removing voxels that have zero value for all subjects. Since the number of samples was not balanced between classes, we used the area under the ROC curve (AUC) as the performance measure to evaluate different methods in this study.

3.2 Evaluation on Subjects with Both MRI and PET

We first evaluated whether the predicted PET data were similar to the true PET data. In the data set used for this study, there were 398 subjects with both MRI and PET images. We randomly sampled 1/2 of these 398 subjects for training the 3-D CNN model. Then the model was used to predict the PET images of the remaining 1/2 subjects. Since we had true PET images for the test subjects, we were able to compare the true and the predicted PET images both visually and quantitatively.

We first visually examined the predicted PET patterns with the ground truth data for each subject. Figure 2 shows the predicted and the ground truth data slice by slice for two subjects. We can observe that the predicted PET patterns are similar to the ground truth. This demonstrates that our deep learning based method can successfully estimate the missing PET data.

To evaluate the proposed data completion method quantitatively, we also compared the classification results based on the true and the predicted PET images. In addition, we report the classification results based on KNN and Zero methods. The AUC values of the three classification tasks based on true PET images and predicted images by three methods are given in Table 1.

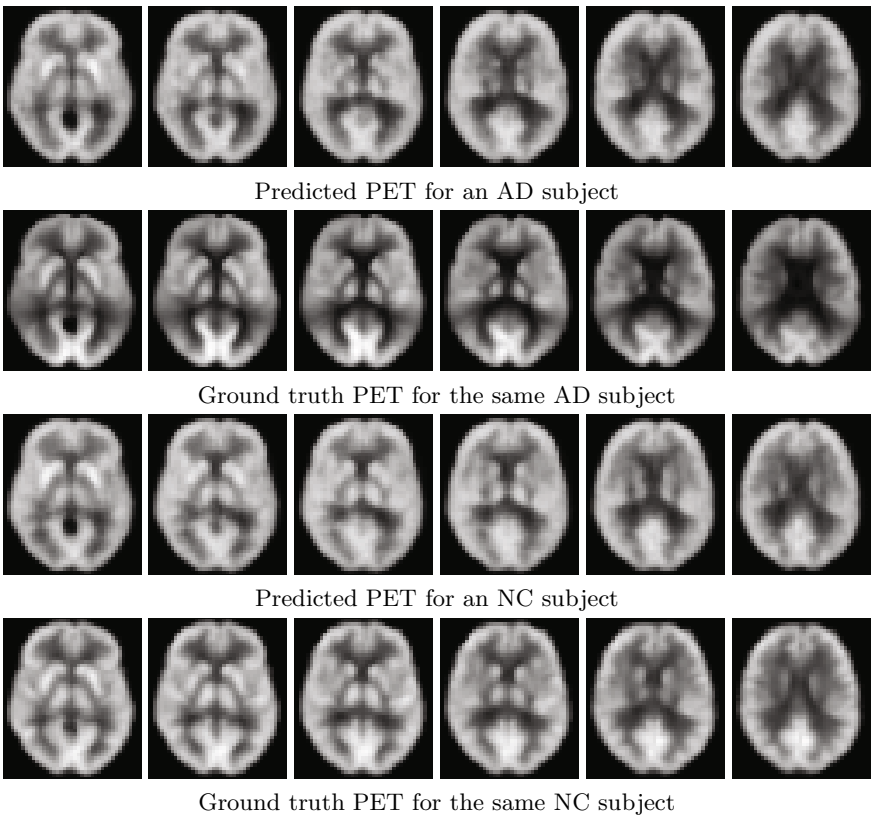


Fig. 2. Comparison of the predicted and the ground truth PET images on two subjects. Each row corresponds to the data (either ground truth or predicted) of one subject, and each column corresponds to slice with the same brain position.

We can observe from these results that the 3-D CNN model outperforms KNN and Zero methods significantly in all three classification tasks. These significant performance differences verify that our deep learning method successfully extracts highly nonlinear relationship between the MRI and PET images. We can also observe that the results of the 3-D CNN model is comparable with those of the true PET images. This demonstrates that our predicted PET images can potentially be used to improve the accuracy of disease diagnosis. Note that the classification performance reported here might be lower than those in the current literature on the ADNI data set because (1) we do not employ advanced feature extraction and classification methods on the true and completed data, and (2) the number of subjects used in the study is relatively small, since we used only these subjects with both MRI and PET.

Table 1. Performance comparison of classification tasks using the true and the predicted PET data. The data set consists of 398 subjects having both MRI and PET images.

Tasks		MCI vs. NC	pMCI vs. sMCI	AD vs. NC
PET	True data	0.7014 ± 0.0212	0.6823 ± 0.0241	0.8982 ± 0.0224
	3-D CNN	0.6947 ± 0.0281	0.6804 ± 0.0267	0.8868 ± 0.0208
	KNN	0.6304 ± 0.0248	0.6278 ± 0.0326	0.7421 ± 0.0282
	Zero	0.6175 ± 0.0213	0.6124 ± 0.0243	0.6928 ± 0.0225

Table 2. Performance comparison of classification tasks using the true and the predicted PET images. All 830 subjects were used in this experiments, where subjects with no PET images were completed using three methods.

Tasks		MCI vs. NC	pMCI vs. sMCI	AD vs. NC
MRI		0.7439 ± 0.0329	0.7168 ± 0.0253	0.9192 ± 0.0188
PET	3-D CNN	0.7305 ± 0.0315	0.7029 ± 0.0245	0.8762 ± 0.0236
	KNN	0.6352 ± 0.0200	0.6133 ± 0.0346	0.7391 ± 0.0304
	Zero	0.6102 ± 0.0268	0.5924 ± 0.0331	0.7028 ± 0.0331
MRI + PET	3-D CNN	0.7621 ± 0.0205	0.7244 ± 0.0241	0.9287 ± 0.0207
	KNN	0.7231 ± 0.0214	0.6813 ± 0.0312	0.7691 ± 0.0213
	Zero	0.7217 ± 0.0290	0.6291 ± 0.0317	0.7003 ± 0.0162

3.3 Evaluation on All Subjects

To further evaluate the effectiveness of our proposed method, we report the prediction performance on all 830 subjects, where 398 subjects have both MRI and PET images, and the remaining 432 subjects have only MRI images. The 3-D CNN and other data completion methods were trained on the 398 subjects, and the trained models were used to complete the PET images of the remaining 432 subjects. The classification performance on all 830 subjects is reported in Table 2. Note that the comparison of classification performance based on true data is not applicable in this experiment, since 432 of 830 subjects did not have PET images.

We can observe that the 3-D CNN model outperforms KNN and Zero methods for all three tasks with three different combinations of PET and MRI modalities. This again demonstrates that the proposed 3-D CNN data completion method is more effective than the competing methods. We can also observe that the performance was improved when the MRI and PET image features were combined. Overall, these experiments yielded insights on the power of the 3-D CNN model in completing missing neuroimaging data, thereby providing practical guidelines for employing multi-modality data even when some data modalities are missing. These results demonstrated that the estimated PET data could be used to improve the accuracy of disease diagnosis.

4 Conclusion and Future Work

We developed a 3-D CNN model for completing and integrating multi-modality neuroimaging data. This model takes one volumetric data modality as input and another modality as output. The nonlinear relationship between different data modalities is captured by a large number of trainable parameters in the network. We applied this model to predict the missing PET patterns from the MRI data. Results showed that the predicted PET data achieved similar classification performance as the true PET images. Additionally, our data completion method significantly outperformed the previous methods.

In this paper, we considered the CNN model for data completion. There are also other deep architectures that achieved promising performance on image-related tasks. It would be interesting to apply other deep models, such as the deep belief networks, for volumetric image data completion. In this work, we employed a CNN model with two hidden layers due to the high computational cost of training. We will explore ways of expediting the computation and design more complicated deep models in the future.

References

1. Weiner, M.W., et al.: The Alzheimers Disease Neuroimaging Initiative: A review of papers published since its inception. Alzheimer's & Dementia 8(suppl. 1), S1–S68 (2012)
2. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
3. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems 25, pp. 1106–1114. Curran Associates, Inc. (2012)
4. Turaga, S.C., Murray, J.F., Jain, V., Roth, F., Helmstaedter, M., Briggman, K., Denk, W., Seung, H.S.: Convolutional networks can learn to generate affinity graphs for image segmentation. Neural Computation 22(2), 511–538 (2010)
5. Jain, V., Seung, S.: Natural image denoising with convolutional networks. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems 21, pp. 769–776. Curran Associates, Inc. (2009)
6. Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Mitosis detection in breast cancer histology images with deep neural networks. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part II. LNCS, vol. 8150, pp. 411–418. Springer, Heidelberg (2013)
7. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(1), 221–231 (2013)
8. Mutch, J., Knoblich, U., Poggio, T.: CNS: a GPU-based framework for simulating cortically-organized networks. Technical Report MIT-CSAIL-TR-2010-013 / CBCL-286, Massachusetts Institute of Technology, Cambridge, MA (February 2010)
9. Yuan, L., Wang, Y., Thompson, P.M., Narayan, V.A., Ye, J.: Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. NeuroImage 61(3), 622–632 (2012)

Human Connectome Module Pattern Detection Using a New Multi-graph MinMax Cut Model

De Wang[1], Yang Wang[2], Feiping Nie[1], Jingwen Yan[2], Weidong Cai[3], Andrew J. Saykin[2], Li Shen[2], and Heng Huang[1,⋆]

[1] Computer Science and Engineering, University of Texas at Arlington, TX, USA
[2] Radiology and Imaging Sciences, Indiana University School of Medicine, IN, USA
[3] BMIT Research Group, School of IT, University of Sydney, Australia

Abstract. Many recent scientific efforts have been devoted to constructing the human connectome using Diffusion Tensor Imaging (DTI) data for understanding the large-scale brain networks that underlie higher-level cognition in human. However, suitable computational network analysis tools are still lacking in human connectome research. To address this problem, we propose a novel multi-graph min-max cut model to detect the consistent network modules from the brain connectivity networks of all studied subjects. A new multi-graph MinMax cut model is introduced to solve this challenging computational neuroscience problem and the efficient optimization algorithm is derived. In the identified connectome module patterns, each network module shows similar connectivity patterns in all subjects, which potentially associate to specific brain functions shared by all subjects. We validate our method by analyzing the weighted fiber connectivity networks. The promising empirical results demonstrate the effectiveness of our method.

1 Introduction

Advent of diffusion MRI technology has made tremendous progress over the last decade [2] and enables us to use Diffusion Tensor Imaging (DTI) for non-invasive in vivo white matter mapping of the human brain by the inference of axonal fiber pathways from local water diffusion [4]. DTI combined with tractography allows the reconstruction of the major fiber bundles in the brain and also permits the mapping of white matter cortico-cortical and cortico-subcortical projections at high spatial resolution. These studies enable the analysis of the human connectome as organizational principle of the central nervous system.

Understanding the structural basis of functional connectivity patterns requires a comprehensive map of structural connection of the human brain, which has been conceptualized as the human connectome [10]. A connectome is a comprehensive description of the network elements and connections that form the brain.

⋆ DW, FN, and HH were supported by IIS-1117965, IIS-1302675, IIS-1344152, DBI-1356628. JY, AS, and LS were supported in part by NSF IIS-1117335, NIH UL1 RR025761, U01 AG024904, NIA RC2 AG036535, NIA R01 AG19771, NIH R01 LM011360, and NIA P30 AG1013318S1.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 313–320, 2014.

Such clear and comprehensive knowledge of anatomical connections lies at the basis of understanding network functions. The connectome can be represented as a large interconnected graph, in which nodes are neuroanatomical regions and synapses are bundles of white matter tracts. The resultant networks exhibit important topological properties such as small-worldness and highly connected hub regions in the posterior medial cortical regions. These studies have accelerated our understandings of human connectome.

Although many network and graph analysis tools have been applied to human connectome studies, most of them focus on analyzing the connectome of each subject individually. How to find the consistent network module patterns (connectome modules) from a group of subjects (*i.e.* a set of regions are connected by similar density of nerve fibers in all subjects) under the same condition (*e.g.* normal or Alzheimer) is important to understand the underlying brain structural and functional mechanisms. The existing research work mainly used the average connectivity networks of all subjects to seek the consistent network modules, however, this straightforward method can easily fail to many conditions. For example, one or two subjects have very strong signals connecting two brain regions, but the rest of subjects have small values on this connectivity. The average connectivity value of all subjects between these two regions can still be large, which indicates a wrong connectivity pattern.

To solve this challenging problem, we propose a novel multi-graph MinMax cut model to identify the consistent network patterns from brain connectivity networks of a group of subjects. Our new approach does the min-max cut on each connectivity network simultaneously. The common connectome patterns are then detected from the dense connected modules. We introduce a new projected gradient optimization algorithm to solve the proposed multi-graph MinMax cut objective. By analyzing the weighted fiber connectivity network from 50 young male adults, we identify six consistent network modules which consistently carry high connectivity among all the subjects. These connectome module patterns potentially associate to the common brain functions shared by all subjects.

2 Methodology

2.1 Consistent Connectivity Patterns

The brain connectome of each subject can be represented as a graph A, in which each node is an ROI (region of interest) in human brain and the weight of each edge is the density of the nerve fibers connecting a pair of nodes. In next section, we will describe the details of brain network construction. Given a group of m subjects under the same condition with n ROIs, we can denote their connectivity networks as $A^1, A^2, \cdots, A^m$, where $A^k \in \Re^{n \times n}$ and A^k_{ij} denotes the connectivity of the i-th ROI and the j-th ROI in the k-th subject, $k = 1, \cdots, m, 1 \leq i, j \leq n$.

It is important to discover the common consistent connectivity patterns, *i.e.* a set of ROIs connected by similar density of nerve fibers in all subjects, which are potentially associated to the underlying brain structural and functional mecha-

nisms shared by the subjects. Thus, our goal is to detect the sub-networks which have similar connectivity structures in all or most $A^1, A^2, \cdots, A^m$.

Although there are many graph cut methods to group nodes in the graph, these approaches only work for single graph and cannot find the common connectome patterns. Thus, we propose a novel multi-graph MinMax cut model to group nodes based on their structures in all connectivity networks.

2.2 Multi-graph MinMax Cut

Given a graph with weight matrix $A \in \Re^{n \times n}$, there are many graph cut methods to group nodes, such as Min Cut, Ratio Cut, Normalized Cut, and MinMax Cut. The MinMax cut can provide the balanced group results to avoid grouping the outlier data together. Thus, MinMax cut is preferred to group nodes in connectome data analysis. However, the traditional MinMax cut only works for single graph. To solve the multiple networks problem, we propose a novel multi-graph MinMax cut model for grouping nodes on multiple graphs simultaneously.

Let $A^v \in \Re^{n \times n}$ denote the v-th network, and D^v are diagonal matrices whose diagonal elements are $\sum_j a_{ij}^v$. When we perform MinMax cut on the v-th network, we can minimize the following spectral relaxed objective [9]:

$$\min_{(Q^v)^T Q^v = I} \sum_{k=1}^{K} \frac{(q_k^v)^T D^v q_k^v}{(q_k^v)^T A^v q_k^v}, \tag{1}$$

where $Q^v = [q_1^v, \cdots, q_K^v] \in \Re^{n \times K}$ is the group indicator matrix for the v-th network and K is the number of groups.

The straightforward way to group ROIs on all networks is to average the corresponding edge weights to build a new "ensemble" network, and perform the MinMax cut on the new network. However, in such method, some networks have very strong signals in local ROIs will dominate the average network and lead to the wrong connectivity patterns. It is ideal to simultaneously perform the MinMax cut on each network and unify their consistent results.

When the multi-graph MinMax algorithm is performed on all networks, the grouping results in different networks should be unique, *i.e.* the group indicator matrices $Q^{(v)}$ of different networks should share the same one. Therefore, in multi-graph MinMax, we force the group assignment matrices to be the same across different networks, that is, the consensus common group indicator matrix $Q \in \Re^{n \times K}$. Our new Multi-Graph MinMax Cut model (MGMMC) is to solve the following objective:

$$\min_{Q^T Q = I} J(Q) = \sum_{v=1}^{m} \sum_{k=1}^{K} \frac{q_k^T D^v q_k}{q_k^T A^v q_k} \tag{2}$$

where m is number of connectivity networks, K is number of clusters. The proposed model is capable of capturing the connectome structures from different

networks, and thus expected to get consistent connectivity patterns. It is difficult to solve the objective in Eq. (2) because of the orthonormality constraints. We will derive our optimization algorithm using the projected gradient descent method.

Taking derivative on $J(Q)$ *w.r.t.* q_k, we get:

$$\frac{\partial J}{\partial q_k} = \sum_{v=1}^{m} \frac{(q_k^T A^v q_k) D^v q_k - (q_k^T D^v q_k) A^v q_k}{(q_k^T A^v q_k)^2} = \sum_{v=1}^{m} \frac{1}{q_k^T A^v q_k} D^v q_k - \frac{q_k^T D^v q_k}{(q_k^T A^v q_k)^2} A^v q_k$$

We denote:

$$\alpha^v = diag(\frac{1}{q_1^T A^v q_1}, \cdots, \frac{1}{q_K^T A^v q_K}),\ \beta^v = diag(\frac{q_1^T D^v q_1}{(q_1^T A^v q_1)^2}, \cdots, \frac{q_K^T D^v q_K}{(q_K^T A^v q_K)^2}),$$

where $diag(x)$ represents a diagonal matrix whose diagonal elements are the elements in vector x.

So Eq. (3) can be rewritten as:

$$\frac{\partial J}{\partial Q} = \sum_{v=1}^{m} D^v Q \alpha^v - A^v Q \beta^v . \tag{3}$$

Because we have the orthonormal constraint $Q^T Q = I$ in objective, we can use the projected gradient descent method to solve this problem. Given Q, we calculate a new variable H by:

$$H = Q - \tau(\sum_{v=1}^{m} D^v Q \alpha^v - A^v Q \beta^v) . \tag{4}$$

When H is fixed, we need to solve the following constrained optimization problem:

$$\min_{Q^T Q = I} \|Q - H\|_F^2 \tag{5}$$

Because

$$\|Q - H\|_F^2 = Tr((Q - H)^T (Q - H)) = Tr(Q^T Q - 2Q^T H + H^T H) , \tag{6}$$

and $Q^T Q = I$, and H is fixed, problem (5) is equivalent to solve the following problem:

$$\max_{Q^T Q = I} Tr(Q^T H) . \tag{7}$$

If the SVD result of H is : $H = U \Sigma V^T$, then the optimal solution of problem (7) can be obtained by:

$$Q = U_K V^T , \tag{8}$$

where U_K is composed of the first K columns of U. Thus,we can iteratively solve H using Eq. (4) and update Q by Eq. (8) till convergence.

3 Human Brain Connectivity Network Construction

In our project, participants included 50 healthy young male adults (age: 24.0 ± 3.2) with no history of neurological or psychiatric disorder. The MRI scans were acquired on a Siemens 3T TIM Trio (Erlangen, Germany) using a 12-channel receive only phased array head coil in combination with a body coil for radio frequency transmission. A SE-EPI DTI sequence was applied using parameters: matrix= 128×128; FOV= 256×256mm; TE/TR=77/8300 ms; 68 transversal slices with 2mm thickness; 48 diffusion directions with gradients b=1000s/mm2, and 8 samplings at b=0. Each session also included a high resolution T1-weighted MP-RAGE imaging as anatomical reference for subsequent parcellation and co-registration.

The DTI data are analyzed in FSL[1]. DTI preprocessing includes correction for motion and eddy current effects in DTI images. The processed DTI images are then output to Diffusion Toolkit (http://trackvis.org/) for fiber tracking, using the streamline tractography algorithm called FACT (fiber assignment by continuous tracking). The FACT algorithm initializes tracks from many seed points and propagates these tracks along the vector of the largest principle axis within each voxel until certain termination criteria are met. In our study, stop angle threshold is set to 35 degree, which means if the angle change between two voxels is greater than 35 degree, the tracking process stops. A spline filtering is then applied to smooth the tracks.

Anatomical parcellation is performed using FreeSurfer 5.1[2] [7,5,6] on the high-resolution T1-weighted anatomical MRI scan acquired with MP-RAGE sequence. The parcellation is an automated operation on each subject to obtain 82 gyral-based ROIs, with 41 cortical ROIs in each hemisphere, one in brainstem. The T1-weighted MRI image is registered to the low resolution b0 image of DTI data using the FLIRT toolbox in FSL, and the warping parameters are applied to the ROIs so that a new set of ROIs in the DTI image space are created. These new ROIs are used for constructing the structural network.

The topological representation of a network is a collection of nodes and edges between pairs of nodes. In constructing the weighted, undirected network, the nodes are chosen to be the 83 registered ROIs obtained from FreeSurfer parcellation. Three different schemes [8,3] are used to define the edge weight as follows: 1) Weighted: The density of the fibers connecting a pair of nodes, which is the number of tracks between two ROIs divided by the mean volume of the two ROIs; 2) Fiber number: the number of tracks between two ROIs; 3) Fiber length: the length of tracks between two ROIs.

4 Experiments and Discussions

4.1 Experiment Setup

We apply our MGMMC model on the 50 connectivity networks. The parameter group number K is set as 10, and the stepsize τ is set as 0.001 for all experiments.

[1] `http://www.fmrib.ox.ac.uk/fsl.html`
[2] `http://surfer.nmr.mgh.harvard.edu/`

We use the normalized connectivity measure of connectome modules to evaluate the density of detected modules:

$$B_{tt} = \frac{\sum_{v=1}^{m} s_{tt}^{v}}{m \cdot n_t^2}, \tag{9}$$

where B_{tt} represents the normalized connectivity of the t-th connectome module, m is the total number of networks used in experiments, n_t is the cardinality of the t-th module C_t (C_t is the set of ROIs contained in the t-th cluster), *i.e.* the number of ROIs in t-th module C_t. $s_{tt}^{v} = \sum_{i \in C_t, j \in C_t} A_{ij}^{v}$ is the connectivity measure of the t-th module in the v-th network.

4.2 Comparison of Connectivity Measures

To demonstrate the effectiveness of our MGMMC model, we compare MGMMC with two methods:
(1) MMC performed on the average network, where $G_{avg} = \frac{\sum_{v=1}^{m} G^v}{m}$.
(2) Multi-Modal Spectral Clustering (MMSC)[1], which integrates data from different modality/view to perform spectral clustering.

The connectivity measure are reported in Table 1. We can conclude that: for all three types of graph (W, LL, NF), the average connectivity measurements of top 6 modules detected by our MGMMC model are greater than that of modules detected using the two comparison methods. This justifies the effectiveness and advantage of our MGMMC model, which considers the connectivity structures in different graphs. T-test is performed to evaluate the significance of difference of the module connectivity. The p values of the T-test for the six pair comparisons (W, W_avg), (W, W_mmsc), (LL, LL_avg), (LL, LL_mmsc), (NF, NF_avg), (NF, NF_mmsc) are 0.14, 0.04, 0.04, 0.008, 0.03, 0.03. Five out of the six p values are less than 0.05, which means the difference of most of the six pair comparisons are significant in all cases except one. We can also see from Table 1 that: the average connectivity measures of detected connectome modules by using weighted network is the best among three types of networks, and the fiber length (LL) network gets the worst connectivity measures. This shows that the weighted network is the best connectivity measurement.

4.3 Visualization of Detected Modules

We visualize the top 6 connectome modules using weight network detected by MGMMC model in Figure 4.2. Only the first 24 subjects are shown due to space limitation. ROIs contained in each connectome module are listed on the left-side in Figure 4.2. We can see that three pairs of modules are almost symmetric except for one ROI in each pair: module 1 and module 2, module 3 and module 6, module 4 and module 5. This shows that each connectome module has its counterpart in the other half brain. In Figure 2, we visualize the location of the top 6 connectome modules in human brain.

Table 1. B_{tt} for top 6 modules with largest connectivity. W, LL and NF denote the results of using our MGMMC model on the graphs of weighted, fiber length, and fiber number, respectively; W_mmc, LL_mmc, and FN_mmc denote the results of MMC using the average graph; W_mmsc, LL_mmsc, and FN_mmsc denote the results using mmsc method.

Module	W	W_mmc	W_mmsc	LL	LL_mmc	LL_mmsc	NF	NF_mmc	NF_mmsc
1	0.1369	0.1250	0.1369	0.0693	0.0649	0.0649	0.0870	0.0829	0.0773
2	0.1003	0.1136	0.0884	0.0649	0.0636	0.0636	0.0818	0.0765	0.0741
3	0.0971	0.0812	0.0867	0.0607	0.0607	0.0545	0.0765	0.0726	0.0735
4	0.0884	0.0780	0.0865	0.0582	0.0522	0.0522	0.0726	0.0658	0.0709
5	0.0867	0.0780	0.0839	0.0522	0.0495	0.0403	0.0640	0.0638	0.0702
6	0.0805	0.0709	0.0768	0.0495	0.0393	0.0386	0.0624	0.0624	0.0574
Average	**0.0983**	0.0911	0.0932	**0.0592**	0.0551	0.0524	**0.0740**	0.0707	0.0706

Fig. 1. The top 6 connectome modules discovered by our MGMMC model. ROIs contained in each connectome module are listed on the left-side. The edge between two nodes denotes there is connection between these two ROIs. Zoom in for clear view.

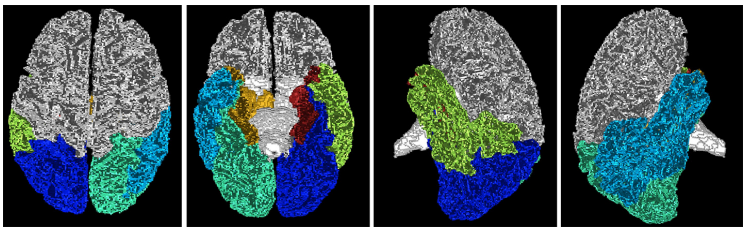


Fig. 2. Location visualization of top 6 connectome modules discovered by MGMMC model from top, bottom, right, left views

5 Conclusion

In this paper, we proposed a novel brain connectivity network analysis method by employing the new multi-graph MinMax cut model to identify the consistent connectivity patterns from multiple subjects. We introduced an efficient algorithm to discover such connectivity patterns that are potentially associated to different brain functions of humans. The clinical DTI data were used to construct the brain connectivity networks to validate our methods. Several important highly connected sub-network modules were detected.

References

1. Cai, X., Nie, F., Huang, H., Kamangar, F.: Heterogeneous image feature integration via multi-modal spectral clustering. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1977–1984. IEEE (2011)
2. Catani, M., Howard, R., Pajevic, S., Jones, D.: Virtual in vivo interactive dissection of white matter fasciculi in the human brain. Neuroimage 17(1), 77–94 (2002)
3. Cheng, H., et al.: Optimization of seed density in dti tractography for structural networks. J. Neurosci. Methods 203(1), 264–272 (2012)
4. Ciccarelli, O., Toosy, A., Parker, G., Wheeler-Kingshott, C., Barker, G., Miller, D., Thompson, A.: Diffusion tractography based group mapping of major white-matter pathways in the human brain. Neuroimage 19(4), 1545–1555 (2003)
5. Dale, A., Fischl, B., Sereno, M.: Cortical surface-based analysis. i. segmentation and surface reconstruction. Neuroimage 9(2), 179–194 (1999)
6. Fischl, B., et al.: Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron. 33(3), 341–355 (2002)
7. Fischl, B., Sereno, M., Dale, A.: Cortical surface-based analysis. ii: Inflation, flattening, and a surface-based coordinate system. Neuroimage 9(2), 195–207 (1999)
8. Hagmann, P., Kurant, M., Gigandet, X., Thiran, P., Wedeen, V.J., Meuli, R., Thiran, J.P.: Mapping human whole-brain structural networks with diffusion MRI. PLoS One 2(7), e597 (2007)
9. Nie, F., Ding, C., Luo, D., Huang, H.: Improved minmax cut graph clustering with nonnegative relaxation. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010, Part II. LNCS (LNAI), vol. 6322, pp. 451–466. Springer, Heidelberg (2010)
10. Sporns, O., Tononi, G., Kotter, R.: The human connectome: a structural description of the human brain. PLoS Comput. Biol. 1(4), e42 (2005)

Max-Margin Based Learning for Discriminative Bayesian Network from Neuroimaging Data

Luping Zhou[1], Lei Wang[1], Lingqiao Liu[2], Philip Ogunbona[1], and Dinggang Shen[3]

[1] University of Wollongong, Australia
[2] University of Adelaide, Australia
[3] University of North Carolina at Chapel Hill, USA

Abstract. Recently, neuroimaging data have been increasingly used to study the causal relationship among brain regions for the understanding and diagnosis of brain diseases. Recent work on sparse Gaussian Bayesian network (SGBN) has shown it as an efficient tool to learn large scale directional brain networks from neuroimaging data. In this paper, we propose a learning approach to constructing SGBNs that are both representative and discriminative for groups in comparison. A max-margin criterion built directly upon the SGBN models is proposed to effectively optimize the classification performance of the SGBNs. The proposed method shows significant improvements over the state-of-the-art works in the discriminative power of SGBNs.

1 Introduction

Neuroimaging techniques have been widely adopted in brain research for analyzing mental diseases, such as the Alzheimer's disease (AD). They could provide more sensitive and consistent assessments for the early diagnosis of disease. Recently, neuroimage analysis is shifting its emphasis from local brain regions to regional interactions (known as brain network) using graph theory [1]. Such analysis is important because brain network change is often a response to damages like mental diseases. Generally a brain network is constructed as follows (Fig. 1). After aligning to a common stereotaxic space, brain images are partitioned into regions of interest (ROI). A brain network is then modeled by a graph with each node corresponding to a brain region and each edge corresponding to the connectivity between regions. Brain "effective connectivity" analysis focuses on the causal relationships between brain regions [1]. The *directionality* is often of interest, because it may disclose the pathways of how one brain region affects the other. Evidence of causal relationship changes has been found in many mental diseases including AD from multiple imaging modalities [2,3], shedding light on discovering novel connectivity-based biomarkers for disease diagnosis.

Early research works in this regard usually require a prior model of connectivity and study only a small number (≤ 10) of brain regions, such as Structural Equation Modeling [4] and Dynamic Causal Modeling [5]. This situation has been improved recently by [2], where a completely data-driven method, denoted as

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 321–328, 2014.

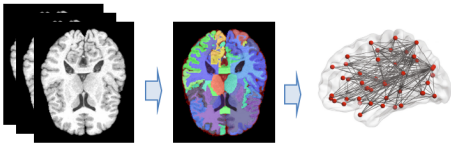


Fig. 1. Illustration of brain network construction from neuroimaging data

H-SGBN in this paper, is proposed to recover sparse Gaussian Bayesian network (SGBN) from more than 40 brain regions in fluorodeoxyglucose PET (FDG-PET) images. It employs the strategy of sparsity constraint to handle large scale Bayesian Network (BN) construction, and circumvents the traditional two-stage procedure for parent set identification in many sparse BN learning methods, achieving a more accurate network recovery [2].

As most BN methods in the literature, H-SGBN is a generative method, which, as pointed out in [6], may ignore the subtle but critical brain structural changes induced by mental diseases. Therefore, a learning approach is proposed in [6], denoted as DL-SGBN, to introduce class discrimination into the SGBN models. DL-SGBN employs Fisher kernel to extract sample-based features from SGBNs, and minimizes a generalization error bound for SVM classifiers with these SGBN-induced features. In that work, the class discrimination is learned by optimizing the classification performance of SVMs, which does not guarantee the equivalent improvement on SGBNs. However, SGBN models are the ultimate goal in such research since they represent the brain connectivity.

In this paper, we propose a new method to learn discriminative SGBN models from neuroimaging data, which overcomes the drawbacks of the state-of-the-art works mentioned above. We propose a max-margin framework to jointly learn two SGBNs, one for each class, for both discrimination and representation. Unlike DL-SGBN in [6], our framework optimizes a criterion directly built upon the classification performance of SGBNs, thus further improves the discriminative power of the models from DL-SGBN (and H-SGBN). Our method is different from the literature of BN classifiers where a *single* BN is learned to represent the *differences* of two classes (in either structure or parameter but not in both) [7,8]. These methods work on discrete variables, while the brain ROI measurements are usually continuous variables whose discretization is often hard to decide. Our experiment shows significant improvement of our proposed method over the state-of-the-art works of H-SGBN and DL-SGBN in terms of the discriminative power of SGBNs. The notations of symbols frequently appearing in this paper are summarized in Table 1.

2 Background

Because this paper is based on sparse Gaussian Bayesian Network (SGBN) model, in the following, we review the fundamentals of SGBN in the original paper [2]. For DL-SGBN, the discriminative learning of SGBN, please refer to [6] for the technical details. We compare with both methods experimentally.

Table 1. Notation

x_i	a random variable
$\mathbf{x}$	a sample of m variables: $\mathbf{x} = [x_1, x_2, \cdots, x_m]^\top$
$\mathbf{X}$	the data matrix of n samples, $\mathbf{X} \in \mathbb{R}^{n \times m}$
$\mathbf{x}_{i,:}$	the i-th row of $\mathbf{X}$, representing a sample
$\mathbf{x}_{:,i}$	the i-th column of $\mathbf{X}$, representing the realization of the random variable x_i on n samples
$\mathbf{W}$	the parameters of a Gaussian Bayesian Network: $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_m]$, $\mathbf{W} \in \mathbb{R}^{m \times m}$
$\mathbf{Pa}_i$	a vector containing the parents of x_i
$\mathbf{PA}_i$	a matrix whose j-th column represents a realization of $\mathbf{Pa}_i$ on the j-th sample.
$\mathbf{G}$	an $m \times m$ matrix for BN: if there is a directed *edge* from x_i to x_j, $\mathbf{G}_{ij} = 1$, otherwise $\mathbf{G}_{ij} = 0$
$\mathbf{P}$	an $m \times m$ matrix for BN: if there is a directed *path* from x_i to x_j, $\mathbf{P}_{ij} = 1$, otherwise $\mathbf{P}_{ij} = 0$

A graph of BN $\mathcal{G}$ expresses the factorization property of a joint distribution $p(\mathbf{x}) = \prod_{i=1,\cdots,m} p(x_i|\mathbf{Pa}_i)$. The conditional probability $p(x_i|\mathbf{Pa}_i)$ is assumed to follow a Gaussian distribution in Gaussian BN (GBN). Each node x_i is regressed over its parent nodes $\mathbf{Pa}_i$: $x_i = \mathbf{w}_i^\top \mathbf{Pa}_i + \varepsilon_i$, where the vector $\mathbf{w}_i$ is the regression coefficients, and $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. A BN is a directed acyclic graph (DAG), i.e., there is no closed path within the graph. Identifying parent sets is critical for BN learning. Traditional methods often consist of two stages: determine candidate parent sets and further prune them by some criteria. A drawback rises that a missing true parent in the first stage will never be recovered. The work in [2] proposed a different approach (H-SGBN) based on sparse GBN (SGBN). In H-SGBN, each node x_i is regressed over all the other nodes, and its parent set is implicitly selected by the regression coefficients $\mathbf{w}_i$ that are estimated by:

$$\min_{\mathbf{W}} \sum_{i=1}^{m} \|\mathbf{x}_{:,i} - \mathbf{PA}_i^\top \mathbf{w}_i\|_2^2 + \lambda_1 \|\mathbf{w}_i\|_1 \tag{1}$$
$$s.t. \quad \mathbf{W}_{ji} \times \mathbf{P}_{ij} = 0, \forall i, j = 1, \cdots, m, \quad i \neq j.$$

All the symbols are defined as in Table 1. A challenge for BN learning is how to enforce the DAG property, i.e., avoiding directed cycles in the graph. A sufficient and necessary condition for being a DAG is proposed in [2], which requires $\mathbf{W}_{ji} \times \mathbf{P}_{ij} = 0$ for all i and j. Note that $\mathbf{P}_{ij}$ is an implicit function of $\mathbf{W}_{ji}$. H-SGBN has been shown to outperform the conventional two-stage methods with higher accuracy for the network edge recovery in [2].

3 Our Proposed Method

As a generative model, BN models the density of the data, revealing how the data could be generated through an underlying process. This is desirable in

the exploratory research of brain, where discovering new knowledges about the brain and the mental diseases is critical. When used for classification, a BN is trained for each class independently and a sample is categorized to the class that produces the higher probability. However, the BNs individually trained by each class may ignore some subtle but critical network differences that distinguish two classes. Since we usually have access to both classes in comparison (e.g., AD and normal control), it is argued in [6] that the parameters of the two SGBNs, one for each class, should be learned from the two classes jointly in order to retain the essential discrimination. Therefore, a joint learning method DL-SGBN is proposed in [6], which introduces group discrimination into SGBNs by optimizing the performance of SVM classifiers with SGBN-induced features. Although this leads to a relatively simple optimization problem, optimizing the performance of SVMs does not necessarily equal to optimizing the discrimination of SGBNs that represent the brain networks. We believe that, the discrimination of SGBNs can be further improved if we *directly* optimize their (instead of SVMs') classification performance. Therefore we propose a new learning framework based on max-margin formulation directly built on SGBNs. We call our method MM-SGBN.

For binary classification, maximizing the minimum margin between two classes can be obtained by maximizing the minimum conditional likelihood ratio (MCLR):

$$\mathrm{MCLR}(\mathbf{W}) = \min_{i=1}^{n} \frac{P(y_i|\mathbf{x}_i, \mathbf{W}_{y_i})}{P(\bar{y}_i|\mathbf{x}_i, \mathbf{W}_{\bar{y}_i})},$$

where n is the number of samples. Without loss of generality, y_i and $\bar{y}_i \in \{-1, 1\}$, representing the true and false labels for the i-th sample, respectively. The parameter $\mathbf{W}_{y_i} = \mathbf{W}_1$ if $y_i = 1$, or $\mathbf{W}_{y_i} = \mathbf{W}_2$ if $y_i = -1$. We can see that MCLR identifies the most confusing sample whose probability of the true class assignment is close to or even less than that of the false class assignment. Hence, maximizing MCLR targets the maximal separation of the most confusing samples in the two classes. It is not difficult to see that MCLR can naturally handle multi-class case when replacing the denominator by the maximal probability induced by all false class assignments. Taking log-likelihood of MCLR, we have

$$\log\ \mathrm{MCLR}(\mathbf{W}) = \min_{i=1}^{n} (\log\ p(\mathbf{x}_i|y_i, \mathbf{W}_{y_i}) - \log\ p(\mathbf{x}_i|\bar{y}_i, \mathbf{W}_{\bar{y}_i})) + const, \quad (2)$$

which can be shown as a quadratic function of $\mathbf{W}$ in the case of SGBN. In order to maximize MCLR, we require the log-likelihood difference in Eqn. (2) larger than a margin for all samples and maximize the margin. To deal with hard separations, we employ a soft margin formulation as follows.

$$\min_{\mathbf{W}_1, \mathbf{W}_2, \xi_i, r} \quad \lambda \sum_{i=1}^{n} \xi_i - r \quad (3)$$

$$s.t. \quad y_i \left(\mathcal{L}(\mathbf{W}_1, \mathbf{x}_i) - \mathcal{L}(\mathbf{W}_2, \mathbf{x}_i)\right) \geq r - \xi_i, \quad \forall i \quad (3a)$$

$$\xi_i \geq 0, \quad r \geq 0, \quad (3b)$$

$$f(\mathbf{X}_1, \mathbf{W}_1) \leq T_1, \quad f(\mathbf{X}_2, \mathbf{W}_2) \leq T_2 \quad (3c)$$

$$\mathbf{W}_1 \in DAG, \quad \mathbf{W}_2 \in DAG \quad (3d)$$

Algorithm 1. MM-SGBN: Discriminative Learning

Input: data $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{n\times m}$, label $\mathbf{y} \in \mathbb{R}^{n\times 1}$

1. Obtain the initial solution for Eqn. (3):
   Get initial $\mathbf{W}^{(0)} = [\mathbf{W}_1^{(0)}, \mathbf{W}_2^{(0)}]$ by Eqn. (1);
   Get initial $r^{(0)}$ and $\epsilon_i^{(0)}$ by solving Eqn. (3) with only the two constraints (3a) and (3b) and a fixed $\mathbf{W} = \mathbf{W}^{(0)}$.
2. Select a subsets of parameters $(\mathbf{W}_{i,j})$ that satisfy:
   i) the gradient (change) of SGBN model at these parameters are highly correlated with the class label, and ii) the corresponding edges present in the graph.
3. Optimize the parameters of the selected nodes by Eqn.(3).

Eqn. (3) has three components addressing class separation (3a), model representation (3c) and DAG property (3d), respectively.

The constraints in (3a) enforce the likelihood of $\mathbf{x}_i$ to its true class larger than that to its false class by a margin r. The variable ξ_i is the slack variables indicating the intrusion of the margin. The function $\mathcal{L}(\cdot)$ denotes the log-likelihood:

$$\mathcal{L}(\mathbf{W}, \mathbf{x}) = \sum_{i=1}^{m} \frac{-(x_i - \mathbf{P}\mathbf{a}_i^\top \mathbf{w}_i)^2}{2\sigma_i^2} - \log(2\pi\sqrt{\sigma_i}).$$

The constraints in (3c) control the fitting errors to maintain reasonable representation. Adding these constraints also avoids the scaling problem of $\mathbf{W}$. The function $f(\cdot)$ measures the squared fitting errors of the corresponding SGBNs for the data $\mathbf{X}_1$ and $\mathbf{X}_2$ from the two classes. It is defined as

$$f(\mathbf{X}, \mathbf{W}) = \sum_{i=1}^{m} \|\mathbf{x}_{:,i} - \mathbf{P}\mathbf{A}_i^\top \mathbf{w}_i\|_2^2.$$

The parameters of T_1 and T_2 are application dependent and predefined by users to control how much representation could be sacrificed for discrimination.

The constraints in (3d) are the DAG constraint proposed in Eqn. (1), i.e., $\mathbf{W}_{1\{ji\}} \times \mathbf{P}_{1\{ij\}} = 0,\ \mathbf{W}_{2\{ji\}} \times \mathbf{P}_{2\{ij\}} = 0, \forall i, j = 1, \cdots, m,\quad i \neq j$. By these constraints, we enforce the validity of both graphs.

The optimization in Eqn. (3) is quadratic programming, which can be solved iteratively by fmincon-SQP (sequential quadratic programming) in Matlab. The details are given in Algorithm 1.

Our method differs from the conventional BN classifiers [7,8] that solely focus on classification. In those methods, only one BN is learned to merely represent the "difference" of the two classes. They no longer model the individual class as our method does, and hence are less interpretative. Moreover, they cannot handle the continuous variables of brain imaging measures, and inherit the drawbacks of the traditional two-stage methods. In practice, learning the whole set of SGBN parameters could become unreliable when the training samples are insufficient. Therefore, we follow the line in [6] to optimize only a selected subset of parameters. Note that this does not introduce the same problem as the

traditional two-stage methods. It is just an engineering trick to handle small sample size problem and becomes unnecessary when sufficient training data are available. In contrast, identifying the candidate-parent sets is an indispensable step in two-stage methods to obtain computationally tractable solutions.

4 Experiment

We evaluate our proposed MM-SGBN against the single class method H-SGBN from [2] and the discriminative learning method DL-SGBN from [6]. For comparison, following [6], we apply all methods on the publicly accessible ADNI[1] database to analyze brain effective connectivity for AD. Three data sets are used from two imaging modalities of MRI and FDG-PET downloaded from ADNI.

MRI data set includes 120 T1-weighted MR images belonging to 50 mild cognitive impairment (MCI) patients and 70 normal controls (NC). These images are preprocessed by the typical procedure of intensity correction, skull stripping, and cerebellum removal. They are segmented into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) using the standard FSL[2] package, and parcellate them into Regions of Interest (ROI) based on an ROI atlas after spatial normalization. The GM volumes of each ROI are used as network nodes. Forty ROIs similar to [6] are used[3], mainly in the temporal lobe and around.
PET data set includes 103 FDG-PET images (and their corresponding MR images) of 51 AD patients and 52 NC. The MR images belonging to different subjects are co-registered and partitioned into ROIs as mentioned above. The ROI partitions are copied onto their corresponding PET images by a rigid transformation. The average tracer uptakes within each ROI in PET images are used as network nodes. Forty discriminative ROIs to AD are used.
MRI-II data set is similar to the MRI data set but using 40 different ROIs covering the typical brain regions spread over the frontal, parietal, occipital and temporal lobes.

We test how the learning process improves the discriminative power of the individual SGBNs estimated by each class. The individual SGBNs are obtained by H-SGBN. We test two methods for discriminative learning: our max-margin-based method MM-SGBN and DL-SGBN in [6]. In order to maintain representation capability, we allow maximal 1% additional squared fitting errors (that is, $T_i = 1.01 \times T_{i0},\ (i = 1, 2)$, where T_{i0} is the squared fitting error of the initial solution) to be introduced during the learning process. To classify a test sample, we compare the values of its likelihood and assign the sample to the class with a higher likelihood. The test accuracies are averaged over the 50 randomly partitioned training-test groups and presented in Table. 2. Paired t-tests (two-tailed) are also conducted to examine the statistical significance of the results.

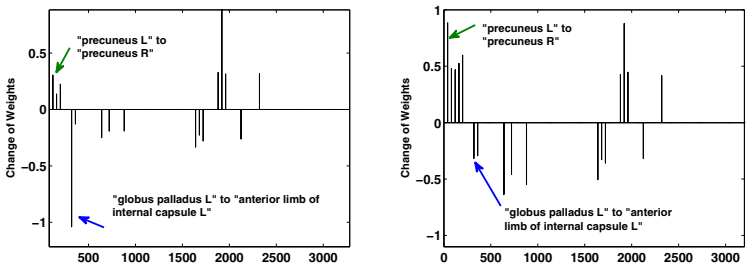
[1] `http://www.adni-info.org/`
[2] `http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/`
[3] Forty ROIs are used to be comparable to that in [2,6].

Table 2. Test Classification Accuracy Averaged over 50 Training-Test Groups (left) and p-values of Paired t-tests (right)

	Accuracy			p-value		
	H-SGBN (%)	DL-SGBN (%)	MM-SGBN (%)	H-SGBN vs. DL-SGBN	H-SGBN vs. MM-SGBN	DL-SGBN vs. MM-SGBN
MRI	66.08	72.92	**76.25**	7e-7	0	1e-4
PET	61.47	66.74	**69.92**	4e-4	0	5e-6
MRI-II	57.08	63.92	**67.17**	7e-6	0	3e-3

From the results we observe that: **i)** Both DL-SGBN and MM-SGBN can greatly improve the discriminative power of the SGBNs estimated from individual classes by H-SGBN. DL-SGBN increases the test accuracy by 6.8% for MRI, 5.3% for PET and 6.8% for MRI-II. MM-SGBN increases the test accuracy by 10.2% for MRI, 8.5% for PET and 10.1% for MRI-II. These improvements are all statistically significant as shown by the very small p-values. This indicates the effectiveness of jointly learning two classes. **ii)** Our proposed MM-SGBN generates the best classification accuracies over all the data sets, which also further improves the classification accuracy of the DL-SGBN by 3.4% for MRI, 3.2% for PET and 3.3% for MRI-II. These improvements are all statistically significant. The advantages of MM-SGBN over DL-SGBN come from directly optimizing the discriminative power of SGBNs, instead of getting indirect help from optimizing the performance of SVM on SGBN-induced features. **iii)** Remind that these improvements on discrimination are achieved with no more than 1% increase of squared fitting errors, as explicitly controlled via the user-defined parameters T_1 and T_2. Note that the rate of 1% is application dependent. More tolerance of fitting errors can potentially bring more discrimination. When we relax fitting error to 10%, another 3% increase of test accuracy could be further achieved.

An example of 18 edge weight changes learned by DL-SGBN and MM-SGBN on PET data is given in Fig. 2, where the SGBN networks from two classes are vectorized and concatenated as x-axis. As shown, both methods learn similar



(a)DL-SGBN (b) MM-SGBN

Fig. 2. Change of edge weights learned by DL-SGBN and MM-SGBN

discriminative patterns despite of using different learning criteria. However, MM-SGBN significantly increases the positive weight of the edge from "precuneus L" to "precuneus R", and reduces the negative weight from "globus palladus L to "anterior limb of internal capsule L". Such differences may lead to the superior performance of MM-SGBN on this dataset and are worthy of further research.

5 Conclusion

In this paper, we propose a max-margin framework directly built on SGBN models to learn causal relationship of brain regions from neuroimaging data. Compared with the state-of-the-art, our method significantly improves the discrimination of the obtained SGBNs, as well as maintaining good representation capacity of the SGBN models.

References

1. Bressler, S., Menon, V.: Large-scale brain networks in cognition: emerging methods and principles. Trends Cogn. Sci. 14(6), 227–290 (2010)
2. Huang, S., Li, J., Ye, J., Fleisher, A., Chen, K., Wu, T., Reiman, E.: A sparse structure learning algorithm for gaussian bayesian network identification from high-dimensional data. IEEE TPAMI 35(6), 1328–1342 (2013)
3. Li, X., Coyle, D., Maguire, L., Watson, D., McGinnity, T.: Gray matter concentration and effective connectivity changes in alzheimer's disease: A longitudinal structural mri study. Neuroradiology 53(10), 733–748 (2011)
4. Kim, J., Zhu, W., Chang, L., Bentler, P., Ernst, T.: Unified structural equation modeling approach for the analysis of multisubject, multivariate functional mri data. Human Brain Mapping 28, 85–93 (2007)
5. Friston, K., Harrison, L., Penney, W.: Dynamic causal modeling. Neuroimage 19, 1273–1302 (2003)
6. Zhou, L., Wang, L., Liu, L., Ogunbona, P., Shen, D.: Discriminative brain effective connectivity analysis for alzheimers disease: A kernel learning approach upon sparse gaussian bayesian network. In: CVPR, pp. 2243–2250 (2013)
7. Pernkopf, F., Bilmes, J.: Efficient heuristics for discriminative structure learning of bayesian network classifiers. J. Mach. Learn. Res. 11, 2323–2360 (2010)
8. Guo, Y., Wilkinson, D., Schuurmans, D.: Maximum margin bayesian networks. In: UAI, pp. 233–242. AUAI Press (2005)

A Novel Structure-Aware Sparse Learning Algorithm for Brain Imaging Genetics

Lei Du[1,⋆] Jingwen Yan[1,2,⋆], Sungeun Kim[1], Shannon L. Risacher[1], Heng Huang[3], Mark Inlow[4], Jason H. Moore[5], Andrew J. Saykin[1], and Li Shen[1,2,⋆⋆], for the Alzheimer's Disease Neuroimaging Initiative[⋆⋆⋆]

[1] Radiology and Imaging Sciences, Indiana University School of Medicine, IN, USA
[2] School of Informatics and Computing, Indiana University Indianapolis, IN, USA
[3] Computer Science and Engineering, University of Texas at Arlington, TX, USA
[4] Mathematics, Rose-Hulman Institute of Technology, IN, USA
[5] Genetics, Geisel School of Medicine, Dartmouth College, NH, USA

Abstract. Brain imaging genetics is an emergent research field where the association between genetic variations such as single nucleotide polymorphisms (SNPs) and neuroimaging quantitative traits (QTs) is evaluated. Sparse canonical correlation analysis (SCCA) is a bi-multivariate analysis method that has the potential to reveal complex multi-SNP-multi-QT associations. Most existing SCCA algorithms are designed using the soft threshold strategy, which assumes that the features in the data are independent from each other. This independence assumption usually does not hold in imaging genetic data, and thus inevitably limits the capability of yielding optimal solutions. We propose a novel structure-aware SCCA (denoted as S2CCA) algorithm to not only eliminate the independence assumption for the input data, but also incorporate group-like structure in the model. Empirical comparison with a widely used SCCA implementation, on both simulated and real imaging genetic data, demonstrated that S2CCA could yield improved prediction performance and biologically meaningful findings.

⋆ Equal contribution by L. Du (leidu@iu.edu) and J. Yan (jingyan@umail.iu.edu).
⋆⋆ Correspondence to Li Shen (shenli@iu.edu). This work was supported by NIH R01 LM011360, U01 AG024904 (details available at http://adni.loni.usc.edu), RC2 AG036535, R01 AG19771, P30 AG10133, and NSF IIS-1117335 at IU, by NSF IIS-1117965, IIS-1302675, IIS-1344152, DBI-1356628 at UTA, and by NIH R01 LM011360, R01 LM009012, and R01 LM010098 at Dartmouth.

⋆⋆⋆ Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 329–336, 2014.

1 Introduction

Brain imaging genetics is an emerging research field aiming to identify associations between genetic factors such as single nucleotide polymorphisms (SNPs) and quantitative traits (QTs) extracted from neuroimaging data. While univariate analyses [9] have been widely used to discover single-SNP-single-QT associations, recent studies have also started to perform regression analyses [5] to examine the joint effect of multiple SNPs on one or a few QTs, and bi-multivariate analyses [4,6,10,12] to examine complex multi-SNP-multi-QT associations.

Sparse canonical correlation analysis (SCCA) [7,14] is a bi-multivariate analysis method that has been applied to both real [6] and simulated [4] imaging genetics data, as well as other omics data sets [2,3,7,14]. Most existing SCCA algorithms use the soft threshold strategy for solving the Lasso [7,14] or group Lasso [4,6] regularization terms. However, the soft threshold approach requires the input data $\mathbf{X}$ to have an orthonormal design $\mathbf{X^T X} = \mathbf{I}$ (see Section 10 in [11]), meaning that the features in the data should be independent from each other. However, for neuroimaging and genetics data, correlation usually exists among regions of interest (ROIs) in the brain and among linkage disequilibirum (LD) blocks in the genome. Simply treating the covariance of the input data as an identity or diagonal matrix will inevitably limit the capability of identifying meaningful imaging genetic associations.

One possible solution to address this issue is to orthogonalize the input data by performing principal component analysis (PCA) before running SCCA. However, we aim to identify relevant imaging and genetic markers, and thus prefer a sparse model. The combined PCA and SCCA strategy cannot achieve this goal, since PCA loadings on the original imaging and genetic markers are non-sparse.

To overcome this limitation, in this paper, we propose a novel structure-aware SCCA (denoted as S2CCA) algorithm for brain imaging genetics applications to achieve the following two goals: (1) our algorithm is not based on the soft threshold framework and eliminates the independence assumption for the input data; (2) our model can incorporate group-like structure (e.g., voxels in an ROI, or SNPs in an LD block) to yield more stable and biologically more meaningful results than conventional SCCA model. We perform an empirical comparison between the proposed S2CCA algorithm and a widely used SCCA implementation in the PMD software package (http://cran.r-project.org/web/packages/PMA/) [14] using both simulated and real imaging genetic data. The empirical results demonstrate that the proposed S2CCA algorithm can yield improved prediction performance and biologically meaningful findings.

2 Structure-aware SCCA (S2CCA)

We denote vectors as boldface lowercase letters and matrices as boldface uppercase ones. For a given matrix $\mathbf{M} = (m_{ij})$, we denote its i-th row and j-th column to $\mathbf{m}^i$ and $\mathbf{m}_j$ respectively. Let $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}^T \subseteq \Re^p$ be the SNP data and $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}^T \subseteq \Re^q$ be the imaging QT data, where n is the number of

participants, p and q are the numbers of SNPs and QTs, respectively. Canonical correlation analysis (CCA) seeks linear combinations of variables in $\mathbf{X}$ and $\mathbf{Y}$ which maximize the correlation between $\mathbf{Xu}$ and $\mathbf{Yv}$:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Yv} \qquad s.t.\ \ \mathbf{u}^T\mathbf{X}^T\mathbf{Xu} = 1, \mathbf{v}^T\mathbf{Y}^T\mathbf{Yv} = 1 \tag{1}$$

where $\mathbf{u}$ and $\mathbf{v}$ are canonical vectors or weights. Two major weaknesses of CCA are that it requires the number of observations n to exceed the combined dimension of $\mathbf{X}$ and $\mathbf{Y}$ and that it produces nonsparse $\mathbf{u}$ and $\mathbf{v}$ which are difficult to interpret. The sparse CCA (SCCA) method removes these weaknesses by maximizing the correlation between $\mathbf{Xu}$ and $\mathbf{Yv}$ subject to the weight vector constraints $P_1(\mathbf{u}) \leq c_1$ and $P_2(\mathbf{v}) \leq c_2$. The penalized matrix decomposition (PMD) toolkit [14] provided a widely used SCCA implementation, where the L_1 penalty $P(A) = \sum_{k=1}^{p} |A(k)|$ was used for both P_1 and P_2. As mentioned earlier, similar to most SCCA methods, PMD employed the soft threshold strategy for solving the L_1 penalty term, which required the input data to have an orthonormal design $\mathbf{X^TX} = \mathbf{I}$ and $\mathbf{Y^TY} = \mathbf{I}$ (see Section 10 in [11]). This independence assumption usually does not hold in imaging genetic data (e.g., correlated voxels in an ROI, correlated SNPs in an LD block), and thus inevitably limits the capability of identifying meaningful imaging genetic associations.

To overcome this limitation, we propose a novel structure-aware SCCA (denoted as S2CCA) algorithm to not only eliminate the independence assumption for the input data, but also incorporate group-like structure in the model. Instead of using L_1, we define a group L_1 constraint on P_1 and P_2 as follows:

$$\begin{aligned} P_1 = ||\mathbf{u}||_G &= \gamma_1 \sum_{k_1=1}^{K_1} \sqrt{\sum_{i \in \pi_{k_1}} u_i^2} = \gamma_1 \sum_{k_1=1}^{K_1} ||\mathbf{u}^{k_1}||_2, \\ P_2 = ||\mathbf{v}||_G &= \gamma_2 \sum_{k_2=1}^{K_2} \sqrt{\sum_{i \in \pi_{k_2}} v_i^2} = \gamma_2 \sum_{k_2=1}^{K_2} ||\mathbf{v}^{k_2}||_2. \end{aligned} \tag{2}$$

In Eq. (2), SNPs are partitioned into K_1 groups $\Pi_1 = \{\pi_{k_1}\}_{k_1=1}^{K_1}$, such that $\{u_i\}_{i=1}^{m_{k_1}} \in \pi_{k_1}$, and m_{k_1} is the number of SNPs in π_{k_1}; and imaging QTs are partitioned into K_2 groups $\Pi_2 = \{\pi_{k_2}\}_{k_2=1}^{K_2}$, such that $\{v_i\}_{i=1}^{m_{k_2}} \in \pi_{k_2}$, and m_{k_2} is the number of QTs in π_{k_2}. $||\cdot||_G$ is the constraint for the group structure. In this work, we partition voxels using AAL ROIs and SNPs using LD blocks.

Now the S2CCA objective function can be formally written as follows:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Yv} - \gamma_1 \sum_{k_1=1}^{K_1} ||\mathbf{u}^{k_1}||_2 - \gamma_2 \sum_{k_2=1}^{K_2} ||\mathbf{v}^{k_2}||_2 \tag{3}$$

$$s.t.\ \ \mathbf{u}^T\mathbf{X}^T\mathbf{Xu} = 1, \mathbf{v}^T\mathbf{Y}^T\mathbf{Yv} = 1,$$

Using Lagrange multipliers, Eq. (3) can be transformed as follows:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Yv} - \gamma_1||\mathbf{u}||_G - \gamma_2||\mathbf{v}||_G - \beta_1||\mathbf{Xu}||_2^2 - \beta_2||\mathbf{Yv}||_2^2 \tag{4}$$

Algorithm 1. Structure-aware SCCA (S2CCA)

Require:
$\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}^T$, $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}^T$
Ensure:
Canonical vectors $\mathbf{u}$ and $\mathbf{v}$.
1: $t = 1$, Initialize $\mathbf{u}_t \in \Re^{p\times 1}$, $\mathbf{v}_t \in \Re^{q\times 1}$;
2: **while** not converged **do**
3: Calculate the block diagonal matrix $\mathbf{D}_{1_t}$, where the k_1-th diagonal is $\frac{1}{2||\mathbf{u}_t^{k_1}||_2}$;
4: $\mathbf{u}_{t+1} = (\beta_1\mathbf{X}^T\mathbf{X} + \gamma_1\mathbf{D}_{1_t})^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{v}_t/2$; Scale $\mathbf{u}_{t+1}$ so that $\mathbf{u}_{t+1}^T\mathbf{X}^T\mathbf{X}\mathbf{u}_{t+1} = 1$;
5: Calculate the block diagonal matrix $\mathbf{D}_{2_t}$, where the k_2-th diagonal is $\frac{1}{2||\mathbf{v}_t^{k_2}||_2}$;
6: $\mathbf{v}_{t+1} = (\beta_2\mathbf{Y}^T\mathbf{Y} + \gamma_2\mathbf{D}_{2_t})^{-1}\mathbf{Y}^T\mathbf{X}\mathbf{u}_{t+1}/2$; Scale $\mathbf{v}_{t+1}$ so that $\mathbf{v}_{t+1}^T\mathbf{Y}^T\mathbf{Y}\mathbf{v}_{t+1} = 1$;
7: $t = t + 1$.
8: **end while**

Taking the derivative about $\mathbf{u}$ and $\mathbf{v}$ and setting them to zero, we have

$$\mathbf{X}^T\mathbf{Y}\mathbf{v}/2 - \gamma_1\mathbf{D}_1\mathbf{u} - \beta_1\mathbf{X}^T\mathbf{X}\mathbf{u} = 0, \tag{5}$$

$$\mathbf{Y}^T\mathbf{X}\mathbf{u}/2 - \gamma_2\mathbf{D}_2\mathbf{v} - \beta_2\mathbf{Y}^T\mathbf{Y}\mathbf{v} = 0, \tag{6}$$

where $\mathbf{D}_1$ is the block diagonal matrix of the k_1-th diagonal block as $\frac{1}{2||\mathbf{u}^{k_1}||_2}$, and $\mathbf{D}_2$ is the block diagonal matrix of the k_2-th diagonal block as $\frac{1}{2||\mathbf{v}^{k_2}||_2}$.

With $\mathbf{v}$ fixed, we can use an approach similar to G-SMuRFS [13] to solve for $\mathbf{u}$. With $\mathbf{u}$ fixed, we can do the same to solve for $\mathbf{v}$. We propose Algorithm 1 to alternatively compute $\mathbf{u}$ and $\mathbf{v}$ until the result converges. We use $\max\{|\delta| \mid \delta \in (\mathbf{u}_{t+1} - \mathbf{u}_t)\} < 10^{-5}$ and $\max\{|\delta| \mid \delta \in (\mathbf{v}_{t+1} - \mathbf{v}_t)\} < 10^{-5}$ as stopping criterion, and nested cross-validation to automatically tune parameters γ_1, γ_2, β_1 and β_2.

3 Experimental Results

3.1 Results on Simulation Data

We first performed a comparative study between S2CCA and PMD using simulated data. We used the following procedure to generate two sets of synthetic data $\mathbf{X}$ and $\mathbf{Y}$, both with $n = 1000$ and $p = q = 50$: 1) We created a random positive definite non-overlapping group structured covariance matrix $\mathbf{M}$. 2) Data set $\mathbf{Y}$ with covariance structure $\mathbf{M}$ was calculated through Cholesky decomposition. 3) We repeated the above two steps to generate another data set $\mathbf{X}$. 4) Canonical loadings $\mathbf{u}$ and $\mathbf{v}$ were set based on the group structures of $\mathbf{X}$ and $\mathbf{Y}$ respectively, where all the variables within the group share the same weights. In this initial study, for simplicity, we selected only one group in $\mathbf{Y}$ to be associated with 4 groups in $\mathbf{X}$. 5) The portion of the specified group in $\mathbf{Y}$ were replaced based on the $\mathbf{u}$, $\mathbf{v}$, $\mathbf{X}$ and the assigned correlation. We generated 7 pairs of $\mathbf{X}$ and $\mathbf{Y}$ with correlations ranging from 0.45 to 0.99. The canonical loadings and group structure remained the same across all the synthetic data sets.

We applied S2CCA and PMD to all seven data sets. The regularization parameters were optimally tuned using a grid search from 10^{-5} to 10^5 through

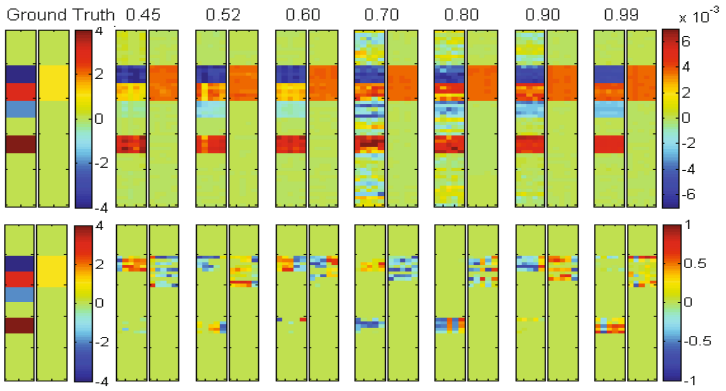


Fig. 1. 5-fold trained weights of **u** and **v**. Ground truth of **u** and **v** are shown in the most left two panels. S2CCA results (top row) and PMD results (bottom row) are shown in the remaining panels, corresponding to true correlation coefficients (CCs) ranging from 0.45 to 0.99. For each panel pair, the five estimated **u** values are shown on the left panel, and the five estimated **v** values are shown on the right panel.

Table 1. Five-fold cross-validation performance on synthetic data: mean±std is shown for estimated correlation coefficients and AUC of the test data using the trained model. P-value of paired t-test between S2CCA and PMD results is also shown.

True	Correlation Coefficient (CC)			Area under ROC (AUC)					
CC	S2CCA	PMD	p	S2CCA:**u**	PMD:**u**	p	S2CCA:**v**	PMD:**v**	p
0.445	0.42±0.05	0.27±0.08	7E-4	1.00±0	0.68±0.02	4E-6	1.00±0	0.84±0.02	4E-5
0.526	0.48±0.04	0.32±0.11	4E-3	1.00±0	0.66±0.01	3E-7	1.00±0	0.87±0.06	3E-3
0.594	0.56±0.07	0.39±0.12	2E-3	1.00±0	0.64±0.01	3E-7	1.00±0	0.81±0.05	7E-4
0.697	0.67±0.01	0.47±0.07	2E-3	0.94±0.02	0.66±0.03	6E-5	1.00±0	0.85±0.04	3E-4
0.814	0.80±0.04	0.49±0.06	7E-5	0.98±0.02	0.63±0.01	1E-6	1.00±0	0.83±0.04	5E-4
0.906	0.90±0.01	0.56±0.06	9E-5	1.00±0	0.66±0.01	4E-7	1.00±0	0.82±0.04	4E-4
1.000	0.99±0.00	0.65±0.04	2E-5	1.00±0	0.66±0.01	3E-7	1.00±0	0.86±0.07	4E-3

nested 5-fold cross-validation. The true and estimated **u** and **v** values are shown in Fig. 1. Due to different normalization strategies, the weights yielded through S2CCA and PMD showed different scales. Yet the overall profile of the estimated **u** and **v** values from S2CCA remained consistent with the ground truth across the entire range of tested correlation strengths (from 0.45 to 0.99), while PMD only identified an incomplete portion of all the signals. Furthermore, we also examined the correlation in the test set computed using the learned CCA models from the training data for both methods. The left part of Table 1 demonstrates that S2CCA outperformed PMD consistently and significantly, and it could accurately reveal the embedded true correlation even in the test data. The right part of Table 1 demonstrates the sensitivity and specificity performance using area under ROC (AUC), where S2CCA also significantly outperformed PMD no matter whether the correlation was weak or strong. From the above results, it can also be observed that S2CCA could identify the correlations and signal locations not only more accurately but also more stably.

Table 2. Participant characteristics

	HC	MCI	AD
Num	304	363	176
Gender(M/F)	111/193	235/128	95/81
Handedness(R/L)	190/14	329/34	166/10
Age (mean±std)	76.07±4.99	74.88±7.37	75.60±7.50
Education (mean±std)	16.15±2.73	15.72±2.30	14.84±3.12

3.2 Results on Real Neuroimaging Genetics Data

S2CCA and PMD were also compared using real neuroimaging and SNP data. The magnetic resonance imaging (MRI) and SNP data were downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. One goal of ADNI has been to test whether serial MRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. For up-to-date information, see www.adni-info.org.

This ADNI study included 176 AD, 363 MCI and 304 healthy control (HC) non-Hispanic Caucasian participants (Table 2). Structural MRI scans were processed with voxel-based morphometry (VBM) in SPM8 [1,8]. Briefly, scans were aligned to a T1-weighted template image, segmented into gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF) maps, normalized to MNI space, and smoothed with an 8mm FWHM kernel. Rather than using ROI summary statistics, in this study we subsampled the whole brain and examined correlations between the voxels (GM density measures) and SNPs. A total of 465 voxels spanning all brain ROIs were extracted. All SNPs within LD block of APOE e4 were extracted from an imputed genetic data set containing only SNPs in Illumina 610Q and/or OmniExpress arrays after basic quality control. As a result, four SNPs (rs429358, rs439401, rs445925, rs534007) from this LD block were included in this study. Using the regression weights derived from the healthy control participants, VBM and genetic measures were pre-adjusted for removing the effects of the baseline age, gender, education, and handedness.

Both S2CCA and PMD were performed on the normalized VBM and SNP measurements. Similar to the previous analysis, 5-fold nested cross-validation was applied to optimally tune the parameters. Table 3 shows 5-fold cross-validation canonical correlation results, indicating that S2CCA significantly and consistently outperformed PMD in terms of identifying high correlations from the training data and replicating those in the testing data. Shown in Fig. 2(a) are the canonical loadings trained from 5-fold cross-validation, suggesting relevant imaging and genetic markers. Although the S2CCA model did not explicitly impose sparsity on individual voxels, it was still able to discover a very small number of relevant ROIs for easy interpretation due to the imposed group sparsity. The strongest imaging signals came from the right hippocampus, which were inversely correlated with APOE e4 allele rs429358. In contrast, despite the flat sparsity design, PMD identified many more ROIs than S2CCA (Fig. 2 (a-b)), making results hard to interpret. In addition, comparing the results from 5 cross-validation trials, S2CCA yielded a more stable and consistent pattern than PMD. It is reassuring that S2CCA

Table 3. Five-fold cross validation canonical correlation results on real data: the CCA models learned from the training data were used to estimate the correlation coefficients between canonical components for both training and testing sets. P-values of paired t-tests were obtained for comparing S2CCA and PMD results.

Correlation coefficients	S2CCA					PMD					p-value
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	
Training	0.28	0.27	0.27	0.27	0.27	0.26	0.26	0.26	0.26	0.24	0.016
Testing	0.21	0.24	0.28	0.23	0.26	0.20	0.21	0.21	0.20	0.24	0.017

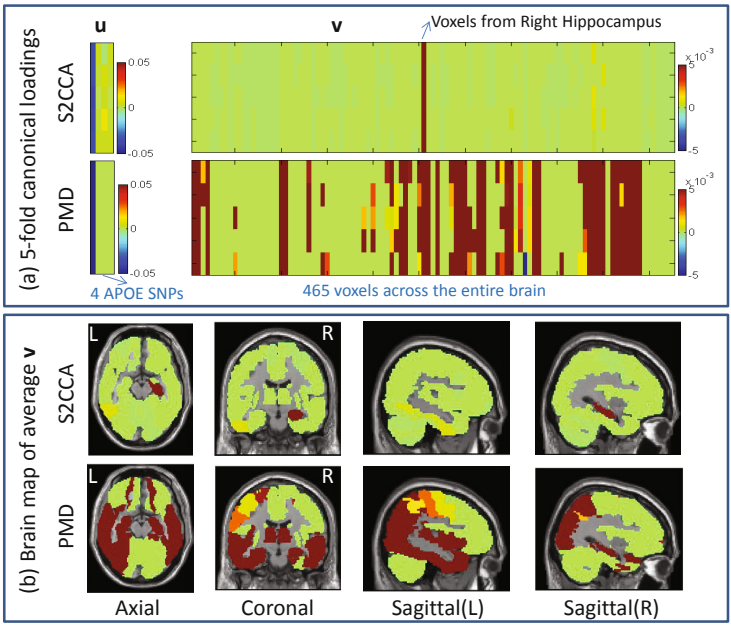


Fig. 2. Comparison of S2CCA and PMD canonical vectors in cross-validation trials: (a) 5-fold canonical loadings of **u** and **v** on 4 APOE SNPs and 465 VBM measures; (b) mapping the average of imaging canonical loadings **v** of 5 cross-validation trials onto the brain

identified a well-known correlation between hippocampal morphometry and APOE in an AD cohort, which shows the promise of S2CCA to correctly identify biologically meaningful imaging genetic associations.

4 Conclusions

Most existing SCCA algorithms (e.g., [4,6,7,12,14]) are designed using the soft threshold strategy, which assumes that the features in the data are independent from each other. This independence assumption usually does not hold in imaging genetic data, and thus limits the capability of yielding optimal results. We have proposed a novel structure-aware sparse canonical correlation analysis (S2CCA) algorithm, which not only removes the above independence assumption, but also

takes into consideration group-like structure in the data. We have compared S2CCA with PMD (a widely used SCCA implementation) on both synthetic data and real imaging genetic data. The promising empirical results demonstrate that S2CCA significantly outperformed PMD in both cases. In addition, S2CCA accurately recovered the true signals from the synthetic data and yielded improved canonical correlation performance and biologically meaningful findings from real data. This study is an initial attempt to remove the feature independence assumption many existing SCCA methods have. Since joint multivariate modeling of imaging genetic data is computationally and statistically challenging, we downsampled our data via a targeted APOE analysis to reduce computational burden and overfitting risk. The S2CCA sparsity was designed to reduce model complexity and further overcome overfitting. Future directions include evaluating S2CCA using more realistic settings and expanding S2CCA to address efficiency and scalability.

References

1. Ashburner, J., Friston, K.J.: Voxel-based morphometry–the methods. Neuroimage 11(6 Pt. 1), 805–821 (2000)
2. Chen, J., Bushman, F.D., et al.: Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. Biostatistics 14(2), 244–258 (2013)
3. Chen, X., Liu, H., Carbonell, J.G.: Structured sparse canonical correlation analysis. In: International Conference on Artificial Intelligence and Statistics (2012)
4. Chi, E., Allen, G., et al.: Imaging genetics via sparse canonical correlation analysis. In: 2013 IEEE 10th Int. Sym. on Biomedical Imaging (ISBI), pp. 740–743 (2013)
5. Hibar, D.P., Kohannim, O., et al.: Multilocus genetic analysis of brain images. Front. Genet. 2, 73 (2011)
6. Lin, D., Calhoun, V.D., Wang, Y.P.: Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. Med. Image Anal. (2013)
7. Parkhomenko, E., Tritchler, D., Beyene, J.: Sparse canonical correlation analysis with application to genomic data integration. Statistical Applications in Genetics and Molecular Biology 8, 1–34 (2009)
8. Risacher, S.L., Saykin, A.J., et al.: Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. Curr. Alzheimer Res. 6(4), 347–361 (2009)
9. Shen, L., Kim, S., et al.: Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. Neuroimage 53(3), 1051–1063 (2010)
10. Sheng, J., Kim, S., et al.: Data synthesis and method evaluation for brain imaging genetics. In: IEEE Int. Sym. on Biomedical Imaging (ISBI), pp. 1202–1205 (2014)
11. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58(1), 267–288 (1996)
12. Vounou, M., Nichols, T.E., Montana, G.: Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. NeuroImage 53(3), 1147–1159 (2010)
13. Wang, H., Nie, F., et al.: Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. Bioinformatics 28(2), 229–237 (2012)
14. Witten, D.M., Tibshirani, R., Hastie, T.: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10(3), 515–534 (2009)

Multi-organ Localization Combining Global-to-Local Regression and Confidence Maps

Romane Gauriau[1,2], Rémi Cuingnet[1], David Lesage[1], and Isabelle Bloch[2]

[1] Philips Research MediSys, Paris, France
[2] Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, Paris, France

Abstract. We propose a method for fast, accurate and robust localization of several organs in medical images. We generalize global-to-local cascades of regression forests [1] to multiple organs. A first regressor encodes global relationships between organs. Subsequent regressors refine the localization of each organ locally and independently for improved accuracy. We introduce *confidence maps*, which incorporate information about both the regression vote distribution and the organ shape through probabilistic atlases. They are used within the cascade itself, to better select the test voxels for the second set of regressors, and to provide richer information than the classical bounding boxes thanks to the shape prior. We demonstrate the robustness and accuracy of our approach through a quantitative evaluation on a large database of 130 CT volumes.

1 Medical Motivation and Overview

With the ever growing size and complexity of $3D$ medical acquisitions, automatic, robust and accurate anatomy localization is of prime interest. First, it enables faster data navigation and visualization of target structures. Secondly, organ localization is a key initialization step for tasks such as segmentation. It is, overall, a crucial component to complex workflows such as treatment follow-up.

General object detection has been deeply studied in computer vision. However algorithms proposed for natural $2D$ scenes are usually not efficient enough (exhaustive scanning of the image) or not even applicable (from $2D$ to $3D$) to the case of anatomical objects. Moreover, medical images often hold specific contextual information, which entails to design specific methods. In the literature we can mostly find three types of approaches for multi-organ localization: classification, regression and atlas-based approaches. As shown in [2], regression-based methods are computationally less expensive (about 25 times less) than atlas-based ones, and then more adapted to clinical contexts. A good overview of the different classification and regression approaches proposed so far can be found in [3]. In this paper we focus on regression-based methods, as their speed and accuracy [2] make them well adapted to clinical contexts. The idea of these approaches is to learn a regression function which relates a voxel and its associated image features to a set of parameters that we want to predict (e.g. organ bounding box). We say that a voxel votes for a set of parameters. The votes from several voxels form a distribution from which we can infer the final result. In [2]

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 337–344, 2014.

the authors have developed a multivariate regression approach where the organs bounding boxes positions are predicted from voxel locations. The authors of [1] enriched this approach by performing a cascade of locally trained regressors. In both works, bounding boxes give a very rough approximation of the target organs (e.g. the liver). In [4] the authors increase the organs parameterization complexity. They perform a joint anatomical landmarks detection and then align shape models. This method gives very good and fast results on shapes such as the lungs or the kidneys, but its application on organs like the stomach or the gallbladder remains questionable, as specific landmarks may be difficult to define.

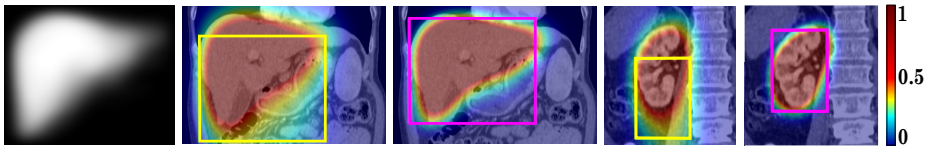
In these methods based on predictions, the authors often use the maximum, the median or the mean of the votes as a final result [2,5]. The vote distributions are not taken into account even though they hold precious information. In this work we intend to make a deeper use of the vote distributions. Following the idea of [6] we also propose to condition the image voxels membership to a global shape prior represented by probabilistic atlases. In Sect. 2 we develop this idea by introducing the concept of *confidence maps* that can be seen as weighted vote distributions associated with shape priors, and propose a fast implementation to compute them. Then we present our global-to-local prediction framework taking benefit of these maps. In Sect. 3 we show different aspects of our contribution: the benefit of performing a greedy parameter optimization and the evaluation of our approach on a large database, thus demonstrating the interest of the *confidence maps* as part of the localization framework and as a result in itself.

2 Methodology

Merging Shape Priors and Vote Distributions with Confidence Maps. In the main works on organ localization with regression, the spatial vote distributions are not fully exploited. However, as shown with the Hough forests [7], vote aggregation can give more information than a single measure of the distribution. For this purpose we introduce the notion of *confidence map*, which encodes the confidence in finding a target organ at a given location. It is built through an aggregation process, making use of both the spatial distribution of regression votes and of organ shape priors through probabilistic atlases.

Probabilistic Atlas. To compute a probabilistic atlas of an organ we first register the binary masks of several samples of this organ. For this purpose let $\{M_i\}_{i\in[\![1,N]\!]}$ be the set of N different cropped binary masks of the organ such that $\forall i\ M_i : \Omega \to \mathbb{R}$ (Ω being the image volume) and where M_r ($r \in [\![1, N]\!]$) is a mask of reference chosen arbitrarily. We transform each mask with a transformation T_i (rigid and anisotropic scaling) in order to scale up the masks M_i to the same size as M_r. The probabilistic atlas A is then computed as an average of these masks, given by $A(x) = \frac{1}{N}\sum_{i=1}^{N} M_i \circ T_i(x)$. Each value $A(x)$ evaluates the probability of a voxel x to belong to the organ (see Fig. 1a).

Confidence Map. Let us consider a regressor $\boldsymbol{R}_\theta$ which, for a test voxel v and its associated image features θ, can give a prediction of the bounding box position



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

Fig. 1. Atlas of the liver (a), localization of the liver and the right kidney: images with confidence map as overlay (images have been cropped) and predicted median box after global step (b,d) and local step (c,e) (best viewed in color)

and scale of an organ with a confidence score α. We parameterize a bounding box by $\mathbf{b} = (b_{min}, b_{max}) \in \mathbb{R}^6$ where b_{min} and b_{max} are its extremal vertices. We denote $X_{\mathbf{b}}$ the set of voxels included in the bounding box with parameterization $\mathbf{b}$. We introduce $\{\mathbf{b}_k\}_{k\in[\![1,K]\!]}$ an ensemble of K votes with confidence scores $\{\alpha_k\}_{k\in[\![1,K]\!]}$. The confidence map C for a given organ is built by localizing and scaling the organ's probabilistic atlas A according to each vote k and by accumulating the result in C with weight α_k. The map C gives a confidence score about the presence of the organ at a given location in the image. The pseudo-code is given in Algo. 1. Some examples of confidence maps are given in Fig. 1(b-e). Figures 1(b,d) show that the maps capture the ambiguity of the vote distribution, as we observe that some voxels were correctly voting for the box upper wall position, an information that the median was not able to capture. The computation of these maps may be expensive. Therefore we propose a fast implementation which considerably reduces the computation time while not degrading significantly the accuracy (see Sect. 3). The idea is to uniformly discretize the space of predicted bounding boxes dimensions. As detailed in Algo. 2, considering we have J discrete values, for each j^{th} discretization value we convolve the volume C'_j, where associated center of votes have been set to their confidence score, with the probabilistic atlas A_j resized to the corresponding sample dimension. The final map C is computed as the sum of each C'_j (we normalize by the maximum). With a uniform discretization on 27 boxes dimensions (3 per spatial dimension, which is good trade-off between speed and precision), the computation is about 30 times faster.

Localizing Organs with Regressors and Confidence Maps. Our approach consists of two steps in a global-to-local fashion. A first regressor, aiming at capturing the spatial relationships between organs, is learned using global information: voxels of the whole image can vote for the positions of all the organs simultaneously. Then, new regressors, dedicated to each organ, are learned using more local information. The benefit of the cascade approach has been already shown in [1]. Here we propose to introduce the use of confidence maps for refining the votes in the cascade of regressors. The selection of voxels which vote in the local step may benefit of the information given by the confidence maps, that is to say the vote distributions and the shape prior. Notice that this method can be applied with any multi-variate regressor. We denote by $o \in [\![1, N_{\text{org}}]\!]$ the

$C \leftarrow 0$;
for $k \leftarrow 1$ **to** K **do**
  $[b_{min}, b_{max}, \alpha] \leftarrow \boldsymbol{R}_\theta[v[k]]$;
  $X_b \leftarrow \texttt{computeXb}(b_{min}, b_{max})$;
  $c \leftarrow \frac{1}{2}(b_{min} + b_{max})$;
  $d \leftarrow \texttt{norm}(b_{min},b_{max})$;
  $A_t \leftarrow \texttt{resizeToBoxDim}(A,d)$;
  foreach $x \in X_b$ **do**
    $C[x] \leftarrow C[x] + \frac{\alpha}{K} A_t(x - c)$;

Algo. 1. Confidence map computation

$C \leftarrow 0$, $d_{max} \leftarrow 0$, $d_{min} \leftarrow$max ;
for $k \leftarrow 1$ **to** K **do**
  $[b_{min}, b_{max}, \alpha[k]] \leftarrow \boldsymbol{R}_\theta[v[k]]$;
  $c \leftarrow \frac{1}{2}(b_{min} + b_{max})$;
  $d[k] \leftarrow \texttt{norm}(b_{min},b_{max})$;
  $d_{max} \leftarrow \texttt{max}(d_{max},d)$;
  $d_{min} \leftarrow \texttt{min}(d_{min},d)$;
$\hat{d} \leftarrow \texttt{quantize}(d_{max},d_{min},J)$;
foreach $\hat{d_j} \in \hat{d}$ **do**
  $indices \leftarrow \texttt{getIndicesOfVotes}(\hat{d_j},d)$;
  $C' \leftarrow 0$;
  foreach $ind \in indices$ **do**
    $C'[ind] \leftarrow C'[ind] + \alpha[ind]$;
  $A_j \leftarrow \texttt{resizeToBoxDim}(A,\hat{d_j})$;
  $C' \leftarrow \texttt{convolve}(C',A_j)$;
  $C \leftarrow C + C'$;

Algo. 2. Fast confidence map computation

indices of the N_{org} organs to localize, and we describe the two steps more precisely hereafter.

Global Step. In the first step, a random subset of K_g voxels $\{v_k\}_{k\in[\![1,K_g]\!]}$ of the image I will vote for the bounding boxes parameters of organ o $\{\mathbf{b}_{k,o}\}_{k\in[\![1,K_g]\!]}$. Votes are performed according to long-range features computed from the image (see Sect. 3 for specific application with regression forest as regressor). These features are chosen to encapsulate global information from the image. Notice that the regressor should be designed such that the relationships between the organs are implicitly embedded during learning (by preserving correlation information between organs positions). Then we compute the confidence map C_o for each organ o using Algo. 1 or 2 and given a probabilistic atlas A_o.

Local Step. The second step aims at improving the previous localization. We relocalize each organ o individually by first computing the binary mask B_o built from the map C_o at a threshold t_g (see Sect. 3). Then we select a random subset of K_l voxels $\{v_k\}_{k\in[\![1,K_l]\!]}$ such that each voxel $B_o(v_k) = 1$. Each voxel v_k votes for the parameters of organ o using a regressor specifically trained for this organ. Contrarily to the previous step this predictor is now learned using short-range features (see Sect. 3) and computed in the vicinity of the organ o thanks to the confidence map so that we give more importance to local information. We then use the votes to compute new and more accurate confidence maps C'_o for each organ. Figures 1(c,d) show the benefit of this local step.

3 Experiments and Results

To validate our approach we propose to localize 6 abdominal organs: the liver, the kidneys, the gallbladder, the spleen and the stomach from various types of 3D CT volumes and using a regression forest as a regressor.

Using the Multivariate Regression Forest as a Regressor. Regression forests have been introduced in [8] and recently popularized in [2] for the purpose of multi-organ localization. This method has proved to be very fast and

quite accurate. As in [2] we use the random forest to regress the parameters **b** of each organ. Random trees are learned from a subset of test voxels from the training images. Each node of each tree contains a $1D$ feature and in each leaf the distributions of the parameters to regress are stored (here multivariate Gaussian distributions). We refer the reader to [2] for more details on the method. This approach is well suited to rough localization. We are able to reinforce its robustness and accuracy using our global-to-local approach with confidence maps. For the first step of the algorithm we use long-range features computed from the image after Gaussian smoothing. As in [1,2], a $1D$ feature corresponds to the difference of mean intensities in two $3D$ patches of random sizes and locations in a certain range. The statistical information stored in the leaf of the random forest regressor allows us to compute confidence scores for each vote. Notice that in practice votes with low confidence scores are discarded. Here we point out that the forest implicitly encodes the organs relationships, as each tree has been built considering the entire set of parameters from all the organs. In the second step we use one random regression forest per organ while using the same kind of features in a more local range (see Sect. 3). Each of these regression forests is learned from test voxels in the vicinity of the corresponding organ thanks to the confidence maps built after the global step. Votes are thus explicitly restricted to a certain neighborhood around each organ.

Database Description and Implementation. Our database is composed of 130 3D CT images coming from 112 patients with diverse medical conditions (healthy and pathological subjects, no organ missing). It includes volumes with varied fields of view, body shapes, resolution and use or not of contrast agents. Slices and inter-slices resolution ranges from 0.5 to 1 mm and from 0.5 to 3 mm, respectively. All the organs have been manually segmented in these 130 volumes. The dataset has been split randomly into 50 and 80 volumes for training and testing, respectively. Our method was implemented in C++ and running times are given for a machine with four 2.3 GHz cores and 8 Go RAM.

Off-line Training. To reach the best performances and analyze each aspect of the algorithm, we performed an extensive greedy optimization of the algorithm parameters. Learning one tree with depth 12 takes about 2 minutes. Before learning we decorrelate the data with a whitening transformation.

Greedy Parameters Optimization. For each parameter we performed a 3-fold cross-validation on the training set. The accuracy of the algorithm is measured as the mean distance of the predicted box to the ground truth bounding box. We first initialized every parameter arbitrarily. Then we optimized each parameter one-by-one by grid-search and we replaced its value by the optimized one. Concerning the training parameters we optimized the tree depth, the threshold of the confidence map t_{g1} and the range of the features. We got the best performances with tree depths of 14 and 12 for the global and local steps respectively, $t_{g1} = 0.4$ and the range of features $\Delta_g = [0, 70]^3$, $W_g = [0, 70]^3$ and $\Delta_l = [0, 40]^3$,

$W_l = [0, 40]^3$ for global and local steps respectively (where Δ is the range of distances from the test voxel to the $3D$ patches and W is the range of sizes of the $3D$ patches, all sizes given in millimeters). Regarding the testing parameters, we looked for the best number of votes (with $K_g = 30000$ and $K_l = 10000$) for the final prediction and the best threshold t_{g2} of the confidence map after the global step. Setting $t_{g2} = 0.5$ and keeping respectively 3% and 1% of the votes with best confidence gave the best results.

Learning Phase. The best parameters found after the above optimization were used for the final forests learning on the 50 training volumes. Atlases of each organ were learned on the same dataset. As in [8] we perform bagging for learning all the forests (uniform random draw with replacement). Node optimization was performed with 30 feature tries as a good compromise between speed and accuracy. Hereafter, if not specified, we used 3 trees as it achieves a good compromise between computation time and accuracy.

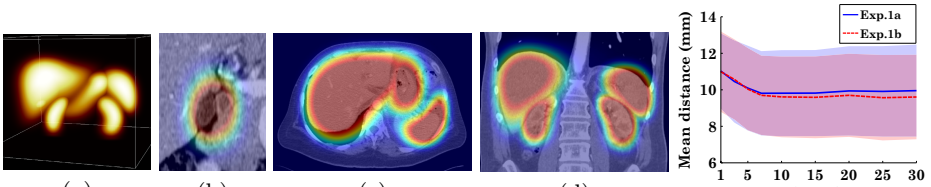
Evaluation of the Localization Approach. The first objective of our evaluation is to show that the confidence maps can be used as a localization result itself, giving more consistent information than the bounding boxes alone. Confidence maps are computed at a 5 mm isotropic spacing. We give some examples of localization results in Fig.2(a-d). An exhaustive visualization of the results can be seen in the **supplementary material**[1]. The lines of Table 1 starting with 'MD' give the mean distance of the thresholded maps contours ($t = 0.5$) to the ground truth contours, using Algo. 1 and 2. The statistics of the results (median and standard deviation) confirm that our method is robust to the variety of test images. We also show that our fast implementation degrades the overall performance very slightly. This allows us to think that our approach can be very useful in various contexts such as segmentation initialization, added to the fact that it runs in about 10 seconds which makes it adapted to clinical applications (code optimization may still improve the computation time). Moreover the confidence maps give much more information than a simple binary mask or contour. For that reason we propose adapted evaluation measures taking into account the fuzziness of the maps. If C denotes a confidence map and B a binary mask of the organ ground truth, then the true positive values are defined as $TP = \sum_{x\in\Omega} B(x)C(x)$, the false negative values as $FN = \sum_{x\in\Omega} B(x)(1 - C(x))$ and the false positive values as $FP = \sum_{x\in\Omega}(1 - B(x))C(x)$. Following the definitions of the sensitivity $S = TP/(TP + FN)$ and the precision $P = TP/(TP + FP)$, we are able to compute the weighted versions of these measures. The corresponding results are reported in Table 1. We notice lower performances for the stomach and the gallbladder, which are challenging organs due to their shape variability. However the figures show that they are still correctly detected.

The second objective of our evaluation comprises two experiments in which we only change the way of selecting the test voxels $\{v_k\}_{k\in[\![1,K_l]\!]}$ for the second local step: exp.1a: from the predicted boxes after the global step, exp.1b: from

[1] `http://perso.telecom-paristech.fr/~gauriau/files/MICCAI14_SupMat.pdf`

Table 1. Results with confidence maps (5 mm isotropic spacing) with Algo.(1, 2) with measures MD: mean distance (mm), S: weighted sensitivity (%), P: weighted precision (%) (mean±standard deviation (median))

		Liver	L. Kidney	R. Kidney	Spleen	Gallbladder	Stomach	**All organs**	Time
Algo.1	MD	9.6 ± 3(9)	5.6 ± 3(5)	6.1 ± 3(6)	7.9 ± 4(7)	9.4 ± 6(7)	13.5 ± 6(12)	8.7 ± 3(8)	
	S	76.8 ± 5(77)	65.7 ± 8(68)	64.6 ± 8(66)	66.3 ± 11(69)	45.2 ± 19(48)	49.7 ± 9(51)	61.4 ± 15(65)	300s
	P	78.8 ± 7(79)	85.2 ± 7(87)	85.1 ± 5(87)	80.8 ± 10(82)	52.4 ± 22(59)	68.3 ± 12(69)	75.1 ± 17(80)	
Algo.2	MD	9.8 ± 3(9)	6.0 ± 4(5)	6.3 ± 3(6)	8.5 ± 4(7)	9.6 ± 4(7)	13.8 ± 6(13)	9.0 ± 3(8)	
	S	75.9 ± 6(77)	64.4 ± 9(66)	63.8 ± 9(66)	64.9 ± 11(67)	44.1 ± 19(46)	49.3 ± 9(50)	60.4 ± 15(64)	10s
	P	78.2 ± 7(79)	84.5 ± 7(86)	85.0 ± 5(87)	80.2 ± 10(82)	52.1 ± 21(57)	67.4 ± 13(68)	74.6 ± 17(80)	



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

Fig. 2. 3*D* MIP of final confidence maps (a), some results with Algo.2 with overlaid confidence maps on the images (b,c,d) and results of exp.1(a,b) (e)

the confidence map using Algo. 2. Figure 2e shows that above 10, the number of trees does not significantly improves the results. Then we keep this number of trees for exp.1(a,b). In Table 2 we report the results of [1,2] and ours after global step and exp.1(a,b). The performances after the global step show that a fine optimization of the parameters helps reaching better results than in [2]. Exp.1a shows the benefit of the cascade approach (additional iterations did not show significant improvements of the results), compared to the single step one. Exp.1b shows the difference between our method and the original cascade approach. The improvement is specially significant for the liver. Moreover Fig. 2 shows that the use of confidence maps tends to rather decrease the bias of the results with an increasing number of trees. We also get similar results to those of [1] for the kidneys, which shows that the cascade approach is scalable and that an increasing number of organs does not degrade the average performance.

Table 2. Box walls mean distances per organ (mean distance (mm) ± standard deviation (median)), per method and per experiment

Method		Liver	L. Kidney	R. Kidney	Spleen	Gallbladder	Stomach	**All organs**	Time(∼)
	[2]*	15.7 ± 15	13.6 ± 13	16.1 ± 16	15.5 ± 15	18.0 ± 15	18.6 ± 16	16.3	4s**
	[1]*	-	7 ± 10(5)	**7 ± 6(6)**	-	-	-	-	3s**
Local	Global	12.5 ± 4(12)	13.6 ± 7(13)	13.8 ± 5(12)	14.3 ± 6(14)	13.9 ± 6(12)	14.3 ± 6(14)	13.9 ± 6(13)	1s
	Exp.1a	11.8 ± 4(11)	6.9 ± 6(5)	**7.2 ± 3(7)**	9.6 ± 7(8)	**9.8 ± 5(9)**	13.6 ± 5(13)	9.8 ± 6(9)	3s
	Exp.1b	**10.5 ± 4(10)**	**6.8 ± 6(5)**	7.3 ± 3(7)	**9.6 ± 6(8)**	10.0 ± 5(8)	**13.5 ± 5(13)**	**9.6 ± 5(8)**	4s

* results are given for other datasets than ours ** times are given for different machines than ours

4 Conclusion

In this article we proposed a fast, robust and accurate method for the localization of multiple organs. We extended the idea of cascade of regressors while introducing the concept of confidence map, which models the vote distributions with the addition of shape prior. We showed that the confidence map, with a proposed fast implementation, can enhance the consistency and accuracy of multi-organ localization with a limited additional computational cost. It is a generic tool with promising potential, which can be used with any type of regressor and which is adaptable to different modalities (e.g. CT, MRI). Moreover its fuzziness property may be useful in many types of clinical applications, such as segmentation (for initialization) or visualization (to target the objects of interest for $3D$ rendering) for instance. Therefore the perspectives are numerous. We also showed that an extensive optimization of the regression parameters significantly improves the localization results. Finally, the consistency and accuracy of our method may still be improved with the use of multiple probabilistic atlases per organ and the regression of the rotation parameters.

Acknowledgments. Work supported in part by an ANRT grant (008512012).

References

1. Cuingnet, R., Prevost, R., Lesage, D., Cohen, L.D., Mory, B., Ardon, R.: Automatic detection and segmentation of kidneys in 3D CT images using random forests. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part III. LNCS, vol. 7512, pp. 66–74. Springer, Heidelberg (2012)
2. Criminisi, A., Robertson, D., Konukoglu, E., Shotton, J., Pathak, S., White, S., Siddiqui, K.: Regression forests for efficient anatomy detection and localization in computed tomography scans. Medical Image Analysis 17(8), 1293–1303 (2013)
3. Zhou, S.: Discriminative anatomy detection: Classification vs regression. Pattern Recognition Letters (in press)
4. Lay, N., Birkbeck, N., Zhang, J., Zhou, S.K.: Rapid multi-organ segmentation using context integration and discriminative models. In: Gee, J.C., Joshi, S., Pohl, K.M., Wells, W.M., Zöllei, L. (eds.) IPMI 2013. LNCS, vol. 7917, pp. 450–462. Springer, Heidelberg (2013)
5. Zhou, S.K., Comaniciu, D.: Shape regression machine. In: Karssemeijer, N., Lelieveldt, B. (eds.) IPMI 2007. LNCS, vol. 4584, pp. 13–25. Springer, Heidelberg (2007)
6. Sun, M., Kohli, P., Shotton, J.: Conditional regression forests for human pose estimation. In: CVPR, pp. 3394–3401 (2012)
7. Gall, J., Lempitsky, V.: Class-specific hough forest for object detection. In: CVPR, pp. 1022–1029 (2009)
8. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)

Inter-Cluster Features for Medical Image Classification

Siyamalan Manivannan, Ruixuan Wang, and Emanuele Trucco

CVIP, School of Computing, University of Dundee, UK
{msiyamalan,ruixuanwang,manueltrucco}@computing.dundee.ac.uk

Abstract. Feature encoding plays an important role for medical image classification. Intra-cluster features such as bag of visual words have been widely used for feature encoding, which are based on the statistical information within each clusters of local features and therefore fail to capture the inter-cluster statistics, such as how the visual words co-occur in images. This paper proposes a new method to choose a subset of cluster pairs based on the idea of Latent Semantic Analysis (LSA) and proposes a new inter-cluster statistics which capture richer information than the traditional co-occurrence information. Since the cluster pairs are selected based on image patches rather than the whole images, the final representation also captures the local structures present in images. Experiments on medical datasets show that explicitly encoding inter-cluster statistics in addition to intra-cluster statistics significantly improves the classification performance, and adding the rich inter-cluster statistics performs better than the frequency based inter-cluster statistics.

1 Introduction

The Bag-of-Words (BoW) approach is widely applied as a feature encoding method for medical [1] as well as natural [2, 3] image classification. In BoW, firstly local features such as SIFT [4] extracted from training images are used to build a dictionary. This dictionary represents a set of visual words (or clusters) which are then used to compute a BoW frequency histogram as a feature vector for any give image. BoW captures the *intra-cluster* statistics of each cluster by just counting the number of local features falling into that cluster (0^{th}-order statistics). On the other hand, *VLAD* [5] and Fisher Vector (FV) [6] represents the intra-cluster information by a rich statistical representation compared to BoW. In VLAD a distance measure between the cluster center and the local features which are assigned to that cluster is used as the intra-cluster information (1^{st}-order statistics). In addition to the 0^{th} and 1^{st} order statistics, FV also considers 2^{nd} order statistics (i.e., variance for each feature component) [6] *within* each cluster. All the above encoding methods (BoW, VLAD and FV) consider that local features extracted from images are independent to each other and none of them captures (1) the *inter-cluster* statistical information (e.g., how two visual words co-occur in each image) and (2) the local structure information of images.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 345–352, 2014.

To capture inter-cluster information, co-occurrences between all pairs of visual words are considered as features for classification [2,3]. However, this leads to a very high-dimensional feature vector. Including inter-cluster features from pairs of clusters which do not have relevant information for classification may decrease classification performance. Recently a mutual information based criterion has been used to select cluster pairs whose co-occurrence information was then used for classification [7]. However, all these methods [2, 3, 7] only consider the dependency between two visual words (first-order co-occurrence) and failed to consider any higher-order dependencies (discussed in section 2). The inter-cluster information in these methods is represented merely as the number of co-occurrence between two clusters. In contrast, we make use of higher-order co-occurrence information to select the informative cluster pairs and encode the inter-cluster features using a richer representation. The contributions of this paper include:

- A new method to select a subset of cluster pairs based on Latent Semantic Analysis (LSA) by considering higher-order co-occurrence of visual words.
- A patch-based method to construct the term-document matrix in the LSA framework, which can capture structural information of objects in images.
- A new inter-cluster feature to capture rich statistical information between selected pairs of clusters, which performs better than co-occurrence frequency.
- Experimental evidence showing that adding inter-cluster statistics (even from a small subset of cluster pairs) improves medical image classification.

2 Inter-Cluster Features

This section focuses on adding inter-cluster statistical information to intra-cluster statistics (e.g., BoW) to represent images. A new method is proposed to choose a subset of cluster pairs by considering the higher-order co-occurrence of visual words within local image regions and introduces an inter-cluster feature which captures rich statistical information between any chosen cluster pairs.

2.1 Selection of Cluster Pairs Based on LSA

Latent Semantic Analysis (LSA) is a well-known technique applied to a wide range of tasks such as search and retrieval [8] and classification [9]. Let $\mathbf{A}$ be a *term-document matrix* with t rows (terms) and d columns (documents), where the element $A(i, j)$ represents the frequency of the occurrence of term i in document j. In image analysis domain, terms correspond to visual words and documents often (but not always, see Section 2.2) correspond to images. In this paper terms and words are used interchangeably. An example of term-document matrix is shown in Figure 2. In LSA, a low-rank (e.g., rank-k) approximation $\mathbf{A}_k$ of matrix $\mathbf{A}$ is obtained by keeping the k largest non-zero singular values in the SVD of $\mathbf{A}$ ($\mathbf{A} = \mathbf{TSD}^{\mathrm{T}}$), i.e., $\mathbf{A}_k = \mathbf{T}_k\mathbf{S}_k\mathbf{D}_k^{\mathrm{T}}$, where the t-by-k matrix $\mathbf{T}_k$, the k-by-k diagonal matrix $\mathbf{S}_k$, and the d-by-k matrix $\mathbf{D}_k$ are respectively the truncated versions of the original matrices $\mathbf{T}$, $\mathbf{S}$, and $\mathbf{D}$. Then the i-th row in

$\mathbf{T}_k\mathbf{S}_k$ can be used to represent the semantic meaning of the i-th term (or word) in the so-called k-dimensional latent semantic space, where noise can be largely suppressed by discarding the smaller singular values in $\mathbf{S}$. Based on such semantic representation of terms, the similarities (correlations) between terms can be captured by the term-term (co-occurrence) matrix, $\mathbf{C}_k = \mathbf{T}_k\mathbf{S}_k(\mathbf{T}_k\mathbf{S}_k)^{\mathsf{T}}$ [10], where each element $C_k(i, j)$ represents the similarity between the i-th and the j-th terms, with higher positive value representing stronger similarity (or positive correlation) between terms and the lower negative value representing stronger anti-similarity (or negative correlation) between terms.

More importantly, it has been shown that term-term matrix $\mathbf{C}_k$ from the truncated matrix $\mathbf{T}_k\mathbf{S}_k$ can additionally capture *higher-order co-occurrence* information (Figure 1) between terms compared to the original co-occurrence matrix (i.e. a matrix where each element (i, j) represents how many times the words i and j co-occur in a document) which is obtained directly from documents [10]. As shown in Figure 1, terms t_1 and t_2, t_2 and t_3, and t_3 and t_4 respectively co-occur in three different documents. With the original co-occurrence matrix, only the first order co-occurrence was captured and therefore the similarity between terms t_1 and t_3 (also t_2 and t_4, and t_1 and t_4) will be zero. But there is a relationship between t_1 and t_3 via t_2. Such higher-order co-occurrence can be captured by the term-term matrix $\mathbf{C}_k$ where the corresponding entries won't be zero.

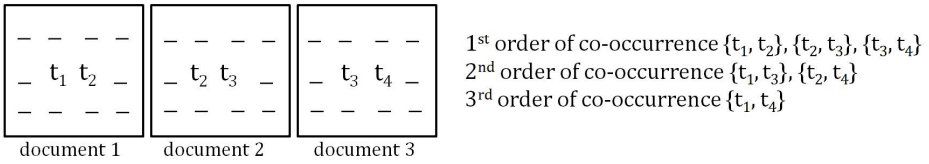


Fig. 1. High-order co-occurrence

We propose to select a subset (say, P percent) of cluster (or term) pairs which have corresponding larger values in the term-term matrix $\mathbf{C}_k$. As explained above, the use of the truncated term-term matrix $\mathbf{C}_k$ instead of the original co-occurrence matrix can help choose the cluster pairs which are semantically similar. In addition, by using a small subset of cluster pairs for inter-cluster feature extraction, richer (in general with higher-dimensional) inter-cluster statistics can be extracted from the selected pairs. Instead, if all the cluster pairs are used for inter-cluster feature extraction as in [2], richer inter-cluster statistics will make feature dimensionality too high to be practically applicable for classifier training.

2.2 Construction of Term-Document Matrix

Note that the truncated term-term matrix $\mathbf{C}_k$ is obtained from the term-document matrix $\mathbf{A}$. To construct $\mathbf{A}$, in general, each image corresponds to one document

and the occurrence of each visual word is counted within the whole image (Figure 2left). However, such term-document matrix construction does not consider any spatial relationship (e.g., far from or close to each other) between the corresponding image regions to any two visual words. As a result, the term-term matrix $\mathbf{C}_k$ won't contain any information about the spatial relationships between any two visual words. In order to make $\mathbf{C}_k$ contain certain spatial relationship between visual words, here we propose to use each image patch (with certain size) as one document (Figure 2right). In this way, the term-term matrix only considers the co-occurrence information between visual words whose corresponding image regions are within the same image patches (therefore close to each other in the image). By selecting word pairs (i, j) whose corresponding absolute values of $C_k(i, j)$ are larger in the patch-based term-term matrix $\mathbf{C}_k$, we expect that the selected highly co-occurred word pairs within image patches (i.e., local image regions) will capture certain structural information of objects in an image, e.g., teeth and nose in radio-graphic images of head often close to each other and therefore more likely appear within an image patch. The statistical information between such cluster (word) pairs may implicitly convey such structural information which cannot be captured within each cluster. What's more, the patch-based term-term matrix $\mathbf{C}_k$ can also capture the larger-scale structural information (if existing) by the higher-order co-occurrence information within $\mathbf{C}_k$, e.g., eye balls with teeth via nose.

2.3 Inter-Cluster Statistics

After selecting a subset of word (or cluster) pairs, we need to extract the inter-cluster information based on these pairs. Let W denote the dictionary which contains N visual words $\{\mathbf{w}_i\}$, and $\mathit{\Pi}$ denote the selected subset of word pairs. Given any image, a number of L local descriptors (e.g., SIFT) $X = \{\mathbf{x}_l, l = 1, \ldots, L\}$ will be extracted from each image patch. Let cluster C_i denote the subset of X such that the nearest visual word for each $\mathbf{x}_l$ in C_i is $\mathbf{w}_i$. We consider the following two measures to respectively capture this inter-cluster statistics:

1. Co-occurrence of visual words: A simple measure of how many times a pair of visual words co-occur locally in each image. Consider an image patch within which visual word $\mathbf{w}_i$ occurs a times and visual word $\mathbf{w}_j$ occurs b times, and the word pair (i, j) is in the selected subset $\mathit{\Pi}$. The co-occurrence statistics $f(i, j)$ of these two visual words inside the image patch will be $f(i, j) = \min(a, b)$.
2. Statistical difference between two clusters: For each cluster C_i, the VLAD descriptor $\mathbf{v}_i$ is first computed as [5] $\mathbf{v}_i = \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{w}_i)$. Then for every word pair (i, j) in $\mathit{\Pi}$, the inter-cluster statistics is computed as $\mathbf{f}(i, j) = ||\frac{\mathbf{v}_i}{\sigma_i} - \frac{\mathbf{v}_j}{\sigma_j}||^2$, where σ_i and σ_j are the standard deviations of the clusters i and j which are computed in the dictionary learning phase by considering all the training features within those clusters. $|| \cdot ||^2$ is a component-wise squared distance measure, and therefore $\mathbf{f}(i, j)$ is a vector and will contain richer statistical information than the scalar co-occurrence value.

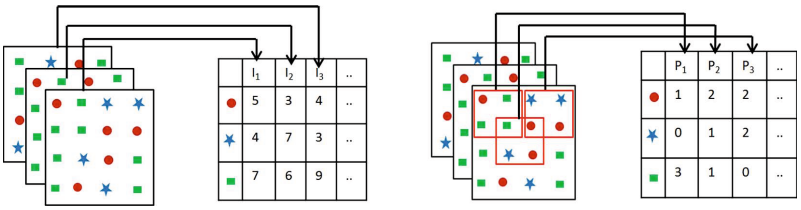


Fig. 2. Term-document matrix obtained from images (left) and patches (right)

2.4 Feature Encoding

Given an image, we encode the image based on both intra-cluster and inter-cluster statistics. First we compute the intra-cluster statistics using the existing approaches such as BoW or VLAD. Then we compute the inter-cluster statistics for image patches in the image as described above. Finally we apply sum pooling over all image patches for the inter-cluster statistics to obtain a feature vector which represents the inter-cluster statistical information for the whole image. The feature vector obtained based on the intra and inter-cluster statistics are normalized individually (we use the power and $L2$ normalizations as in [11]) and concatenated together as the final image descriptor.

3 Experiments

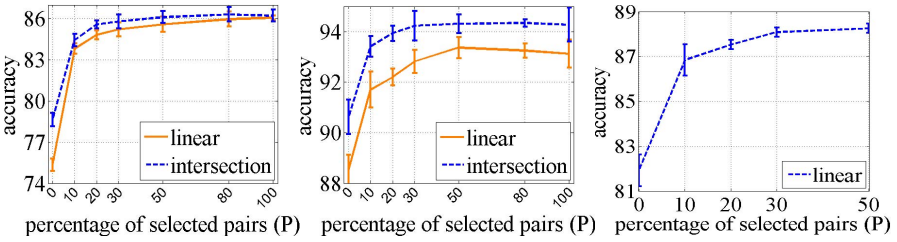
Two medical datasets were used to evaluate the proposed method for cluster pair selection and inter-cluster features. The ICPR HEp-2 cell classification dataset (ICPR[1]) contains $13,596$ gray-scale cell images from 6 classes (homogeneous, speckled, nucleolar, centromere, golgi, and nuclear membrane), with average image size about 70×70 pixels. The Image Retrieval in Medical Applications dataset (IRMA[2]) contains 15,363 anonymous radiographs from 57 classes (of various human body parts), with images resized to be no larger than 300×300. Since the number of images is very unbalanced across IRMA classes, only 20 classes were selected, each of which contains 200 images. We used one-vs-rest multi-class SVM with linear and intersection kernels [12] for classification. SVM parameters were learned using 5-fold cross-validation on the training set. The value of k is chosen such that the $\mathbf{A}_k$ keeps 95% of its column-wise variance. BoW and VLAD features are respectively used as two intra-cluster features based on the local descriptor SIFT, where for each image, dense SIFT descriptors were extracted from each small regions of size 16×16 pixels over a grid with spacing of 4 pixels along both directions, and every 7×7 neighboring regions compose one image patch (i.e., 49 SIFT features in each patch). For ICPR dataset, we applied two-fold cross-validation and report the mean per-class accuracies (MAC) over 5 runs. For the IRMA dataset 30 images per class are selected for training and the rest are used for testing; the averaged MAC over 10 iterations are reported.

[1] `http://i3a2014.unisa.it/`
[2] `http://ganymed.imib.rwth-aachen.de/irma/index_en.php`

3.1 Effect of the Inter-Cluster Features

When using BoW as intra-cluster feature and co-occurrence frequency of visual words as inter-cluster features, Figures 3(a)(b) show that adding inter-cluster features significantly increase the classification performance for both datasets (e.g., around 78% when $P = 0$ vs. 86% when $P > 0$ for ICPR dataset, and around 91% vs. 94% for IRMA dataset, both with dictionary size 200 and using intersection kernel). It also shows that the classification accuracy is not significantly different between selecting 10% (when $P = 10$) and all (when $P = 100$) cluster pairs, which indicates that only a small subset of cluster pairs are sufficient enough to capture the inter-cluster information. Figure 3(a)(b) also show that intersection kernel for intra-cluster feature cannot capture high-order information encoded in inter-cluster features, otherwise adding inter-cluster feature would not improve the accuracy.

Similar findings have been confirmed when using VLAD as the intra-cluster feature and the VLAD-based inter-cluster statistics for the inter-cluster features (Figure 3(c)). By comparing the classification performance from Figures 3(a) and (c), it becomes clear that, even using a smaller dictionary ($N = 32$) and a smaller subset of cluster pairs ($P = 10$ percent), VLAD plus VLAD-based inter-cluster features outperforms the corresponding BoW plus co-occurrences based inter-cluster features, i.e., 86.8% vs. 84.4% for ICPR dataset. Similar finding were found for IRMA dataset (not shown due to limited space). This indicates that both VLAD intra-cluster feature and the VLAD-based inter-cluster feature captures richer statistical information than the BoW intra-cluster feature and the co-occurrence based inter-cluster feature.



(a) ICPR dataset (b) IRMA dataset (c) ICPR dataset

Fig. 3. Effect of the inter-cluster features. $P = 0$ corresponds to intra-cluster feature, and $P > 0$ corresponds to inter-cluster feature plus intra-cluster feature. (a-b) BoW with co-occurrence, (c) VLAD with statistical cluster difference.

To further confirm the effect of inter-cluster features, in Figure 4 left the sizes of the dictionaries are varied and only 20% cluster pairs are chosen based on corresponding dictionaries. It shows a significant performance improvement when adding inter-cluster features, no matter what the dictionary size is. Since adding inter-cluster features for larger dictionaries tremendously increases the dimensionality of the final image representation, in another test, we capture

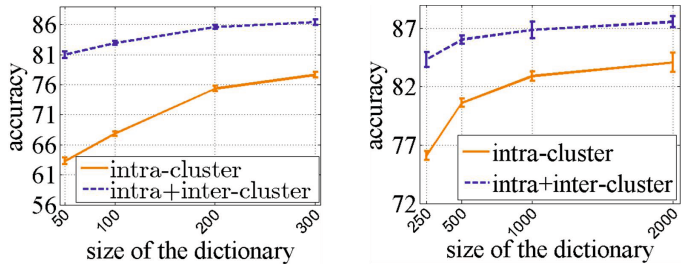


Fig. 4. Classification performance on ICPR dataset with BoW and co-occurrence based inter-cluster features using intersection kernel. See text for more details.

inter-cluster features by considering only 20% pairs from a fixed small dictionary of size 100. Adding these fixed inter-cluster features to the traditional intra-cluster BOW features computed from any larger dictionary still increases the overall performance (Figure 4right). Notice that adding inter-cluster features from a fixed smaller dictionary not only increases the classification accuracy but also reduces the feature dimensionality.

3.2 Patch-Based vs. Image-Based Methods

This test is to compare the performance of patch-based with the image-based cluster pair selection for inter-cluster feature encoding on the IRMA dataset. For both methods, BoW was used as intra-cluster feature and co-occurrence of selected visual words as inter-cluster feature. The dictionary size was fixed to 200 and only 10% of pairs are selected to encode inter-cluster features. As expected, patch-based method gives the accuracy of 93.4%, much better than the accuracy 87.0% from image-based method (with standard deviation about 0.7%), supporting that patch-based method helps capture local structural information encoded in inter-cluster features.

3.3 LSA-Based Pair Selection

In this section the LSA-based truncated term-term matrix is compared with the original co-occurrence matrix for pair selection. In this experiment a dataset containing radiographs of heads taken from four different angles collected from the IRMA dataset is considered. This dataset contains 50 images in each of the four classes. By keeping all the other factors (e.g., patch-based term-document construction and VLAD based inter-cluster feature encoding) unchanged, we found that when selecting a small subset ($P = 5$) of pairs for inter-cluster features, the pair selection based on the truncated term-term matrix performs significantly better than based on the original co-occurrence matrix (78.3% vs. 87.2%). This confirms the potential function of LSA-based pair selection in reducing noise and capturing high-order co-occurrence statistics.

3.4 Inter-Cluster Features for Fisher Vector

Some initial experiments with FV was also performed on ICPR dataset to observe the effect of inter-cluster features for FV. Given an image, Fisher vector $\mathbf{F_i}$ for each cluster C_i was computed based on soft-assignments (see [6] for details). The inter-cluster feature between any chosen cluster pair (i, j) was computed as $||\mathbf{F_i} - \mathbf{F_j}||^2$ (component-wise, as for VLAD). With totally 16 clusters being used, accuracy of 85.2% was obtained by FV. In comparison, adding inter-cluster features ($P = 20$) to FV significantly improves the performance to 88.7%.

4 Conclusions

This paper showed that adding inter-cluster features to the intra-cluster features significantly improves medical image classification. A new method was proposed to select a subset of cluster pairs to get the inter-cluster features. Experiments showed that adding rich inter-cluster statistics performs better than only considering the co-occurrence frequency information as the inter-cluster statistical feature. In feature work we plan to select cluster pairs based on discriminative information (i.e., class labels) and add spatial information to final representation.

Acknowledgement. This work is funded by 2011-2016 EU FP7 ERC project "CODIR: colonic disease investigation by robotic hydrocolonoscopy", collaborative between the Universities of Dundee (PI Prof Sir A Cuschieri) and Leeds (PI Prof A Neville).

References

1. Manivannan, S., Wang, R., Trucco, E., Hood, A.: Automatic normal-abnormal video frame classification for colonoscopy. In: ISBI (2013)
2. Yang, Y., Newsam, S.: Spatial pyramid co-occurrence for image classification. In: ICCV, pp. 1465–1472 (2011)
3. Yang, Y., Newsam, S.: Bag-of-visual-words and spatial extensions for land-use classification. In: AGIS, pp. 270–279 (2010)
4. Lowe, D.: Object recognition from local scale-invariant features. In: ICCV (1999)
5. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: CVPR, pp. 3304–3311 (2010)
6. Perronnin, F., Dance, C.R.: Fisher kernels on visual vocabularies for image categorization. In: CVPR, pp. 1–8 (2007)
7. Chen, T., Yap, K.H., Chau, L.P.: From universal bag-of-words to adaptive bag-of-phrases for mobile scene recognition. In: ICIP, pp. 825–828 (2011)
8. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. JASIS 41(6), 391–407 (1990)
9. Zelikovitz, S., Hirsh, H.: Using LSI for text classification in the presence of background text. In: ICIKM, pp. 113–118 (2001)
10. Kontostathis, A., Pottenger, W.M.: A framework for understanding latent semantic indexing performance. In: IPM, pp. 56–73 (2006)
11. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010)
12. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. JMLR (2008)

A Universal and Efficient Method to Compute Maps from Image-Based Prediction Models

Mert R. Sabuncu*

A.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA, USA

Abstract. Discriminative supervised learning algorithms, such as Support Vector Machines, are becoming increasingly popular in biomedical image computing. One of their main uses is to construct image-based prediction models, e.g., for computer aided diagnosis or "mind reading." A major challenge in these applications is the biological interpretation of the machine learning models, which can be arbitrarily complex functions of the input features (e.g., as induced by kernel-based methods). Recent work has proposed several strategies for deriving maps that highlight regions relevant for accurate prediction. Yet most of these methods either rely on strong assumptions about the prediction model (e.g., linearity, sparsity) and/or data (e.g., Gaussianity), or fail to exploit the covariance structure in the data. In this work, we propose a computationally efficient and universal framework for quantifying associations captured by black box machine learning models. Furthermore, our theoretical perspective reveals that examining associations with predictions, in the absence of ground truth labels, can be very informative. We apply the proposed method to machine learning models trained to predict cognitive impairment from structural neuroimaging data. We demonstrate that our approach yields biologically meaningful maps of association.

Keywords: Machine learning, image-based prediction.

1 Introduction

Broadly, there are two approaches in statistical data analysis [1]: generative (i.e., model based or classical) and discriminative (i.e., prediction oriented). While the former offers more interpretable models, the latter can yield more accurate predictions [1]. Over the last decade, discriminative supervised learning models have been widely adopted to analyze biomedical image data, for example to demonstrate that one can accurately predict a clinical diagnosis form imaging measurements, e.g. [2–6]. The main challenge in these studies is the biological interpretation of image-based prediction models.

One way to gain biological insight is to derive maps of association, which have traditionally been obtained via mass-univariate techniques, such as voxel-based

* Supported by NIH NIBIB 1K25EB013649-01 and a BrightFocus grant (AHAF-A2012333). Data used were obtained from ADNI: http://tinyurl.com/ADNI-main.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 353–360, 2014.

morphometry [7]. Motivated by this approach, recent studies have employed various strategies to compute such maps based on multivariate discriminative models, e.g. [8–11]. These techniques attempt to quantify the statistical relevance of voxel-level features with respect to the predicted variable. Several methods to compute feature relevance, or variable importance, have also been proposed in the machine learning literature, e.g. [12–16]. Yet, as we elaborate in the next section, most of these methods suffer from drawbacks being specific to a type of algorithm/model (non-universality).

In this paper, we present a universal and computationally efficient method to examine associations captured by black box machine learning models. Our method does not rely on knowledge about the learning algorithm. Furthermore, we do not make any strong distributional assumptions about the data. In its simplest form, the proposed method simply uses a dataset, on which predictions have been computed. Our theoretical framework demonstrates that, even in the absence of ground truth labels, the associations we quantify can be informative about the underlying biology. We apply the proposed method to compute maps of association from discriminative models trained to predict a clinical or behavioral condition from structural brain magnetic resonance imaging (MRI) scans.

2 Theory

2.1 Motivation

A popular approach for interpreting a linear discriminative model is to examine the weights, e.g. [17]. Yet, as recently pointed out [11], this interpretation can be misleading. Furthermore, it is not clear whether directly examining the estimated model parameters provides any insight about the underlying biology. This is because the model can be arbitrarily inaccurate and thus model parameters alone might provide little information about the target variable. Sampling strategies, e.g. [9, 13, 15] address this issue by randomly perturbing the data and examining the variation in model parameters and/or predictions. This approach, however, typically requires repeatedly running the computationally expensive training step or resorting to approximation strategies. Moreover, it assumes a particular model structure, e.g. linear, sparse, or a tree.

More general methods applicable to any black box prediction model have recently been proposed, e.g. [14, 18]. Yet these techniques often make strong assumptions about the data (e.g. binary or Gaussian) to offer practical solutions. Our goal in this work is to propose a technique for measuring feature relevance that is universal (i.e. applicable to any black box model), computationally efficient, and robust with respect to the data and the algorithm. Furthermore, we would like the proposed method to capture nonlinear relationships, as well. To achieve this, we build on the theoretical framework that was recently used to derive a generalized measure of correlation [19].

2.2 Proposed Feature Relevance Measure and Its Properties

Let's assume a black box predictive model. We will use capital letters to denote random variables. Let $\boldsymbol{X}$ be the input data vector, which is typically high dimensional, e.g., images, P be the prediction produced by the model, and Y be the target variable that we aim to predict, e.g., clinical status. Note P is typically a non-random function of $\boldsymbol{X}$ and we denote the i'th component of $\boldsymbol{X}$ as $\boldsymbol{X}_i$.

A generalized measure of correlation (GMC) between two random variables, say P and Y, can be derived based on the well-known variance decomposition formula [19]:

$$V(Y) = V(E(Y|P)) + E(V(Y|P)), \tag{1}$$

where V denotes (conditional) variance and E denotes (conditional) expectation, defined over appropriate random variables. The first term on the right, $V(E(Y|P))$, can be interpreted as the explained variance of Y by P. Thus the GMC between Y and P, which we denote as $\gamma(Y|P)$, can be defined as:

$$\gamma(Y|P) = \frac{V(E(Y|P))}{V(Y)}. \tag{2}$$

The GMC is a measure of correlation that quantifies both linear and non-linear dependencies [19] and ranges from 0 (no correlation) to 1 (max. correlation).

We expand Eq. 1 by applying another variance decomposition to $V(E(Y|P))$:

$$V(Y) = V(E(E(Y|P)|\boldsymbol{X}_i)) + E(V(E(Y|P)|\boldsymbol{X}_i) + E(V(Y|P)), \tag{3}$$

where $\boldsymbol{X}_i$ is an input variable and $V(E(E(Y|P)|\boldsymbol{X}_i))$ can be viewed as the explained variance of Y by $\boldsymbol{X}_i$, *as captured* by the model's prediction P. Thus, we define the *captured correlation* as:

$$\kappa(Y|P|\boldsymbol{X}_i) = \frac{V(E(E(Y|P)|\boldsymbol{X}_i))}{V(Y)}. \tag{4}$$

Some of the properties of $\kappa(Y|P|\boldsymbol{X}_i)$ are as follows (Proofs of P1-4 are omitted due to space constraints). Note ρ denotes Pearson's correlation.

P1. $0 \leq \kappa(Y|P|\boldsymbol{X}_i) \leq \gamma(Y|P) \leq 1$.
P2. If P and $\boldsymbol{X}_i$ *or* P and Y are independent, then $\kappa(Y|P|\boldsymbol{X}_i) = 0$.
P3. If $\exists f$ s.t. $f(P) = Y$, then $\kappa(Y|P|\boldsymbol{X}_i) = \gamma(Y|\boldsymbol{X}_i)$.
P4. If $\exists g$ s.t. $g(\boldsymbol{X}_i) = P$, then $\kappa(Y|P|\boldsymbol{X}_i) = \gamma(Y|P)$.
P5. If $\rho(\boldsymbol{X}_i, \boldsymbol{X}_j) = \pm 1$, then $\kappa(Y|P|\boldsymbol{X}_i) = \kappa(Y|P|\boldsymbol{X}_j)$.
Proof: If $\exists a \neq 0, b$ s.t. $\boldsymbol{X}_i = a\boldsymbol{X}_j + b$, then, for any Z, $E(Z|\boldsymbol{X}_i) = E(Z|\boldsymbol{X}_j)$. Thus, $V(E(E(Y|P)|\boldsymbol{X}_i)) = V(E(E(Y|P)|\boldsymbol{X}_j))$, where we use $Z \triangleq E(Y|P)$.
P6. If $\rho(E(Y|P), P) = \pm 1$, then $\kappa(Y|P|\boldsymbol{X}_i) = \gamma(Y|P)\gamma(P|\boldsymbol{X}_i)$.
Proof: Define $Z \triangleq E(Y|P)$. If $\exists a \neq 0, b$ s.t. $Z = aP + b$, then $a^2V(P) = V(Z)$ and $\frac{V(E(Z|\boldsymbol{X}_i))}{a^2} = V(E(P|\boldsymbol{X}_i))$. Then $\kappa(Y|P|\boldsymbol{X}_i) = \frac{V(E(Z|\boldsymbol{X}_i))}{V(Y)}\frac{V(P)a^2}{V(P)a^2} = \frac{V(E(Z|\boldsymbol{X}_i))}{a^2V(Y)}\frac{a^2V(P)}{V(P)} = \frac{V(E(P|\boldsymbol{X}_i))}{V(Y)}\frac{V(Z)}{V(P)} = \gamma(Y|P)\gamma(P|\boldsymbol{X}_i)$.

The first five properties summarize the general behavior of κ as a dependency measure. For example, it is zero if the model's prediction is independent of the variable $\boldsymbol{X}_i$. If the model is perfectly accurate, κ reduces to the GMC between Y and $\boldsymbol{X}_i$. Moreover, as **P5** suggests, the captured correlation is indifferent to whether a variable is directly used in the prediction or correlated alternatives are. Thanks to this property, captured correlation will not highlight an arbitrary subset among correlated variables, the way sparse models do.

P6 is a particularly interesting property, which states that under a specific condition, the captured correlation is proportional to the GMC between the prediction P and input variable $\boldsymbol{X}_i$. We note that, in fact $E(Y|P) = P$ is a common modeling assumption that seems to hold in many practical problems. For example, many regression models (where Y is continuous), assume a zero-mean independent additive Gaussian noise model. Or, in binary classification, P can be the probability of class 1. In both examples, these models imply $E(Y|P) = P$ and thus $\kappa(Y|P|\boldsymbol{X}_i) \propto \gamma(P|\boldsymbol{X}_i)$. In this case, the ranking of variables with respect to their captured correlations is the same as their ranking with respect to their GMC with the prediction. This is a critical observation. It suggests that, in the absence of ground truth data, examining the associations between input variables and the model's predictions can be informative about the relationships with the (ground truth) target variable.

2.3 A Non-parametric Estimator

We propose to employ a non-parametric strategy, which relies on the mild distributional assumption of finite first and second order moments, to estimate the correlation measures κ and γ. Here, we assume that we have access to N independent samples of (X, P, Y), where for notational simplicity we have replaced $\boldsymbol{X}_i$ with X. We denote these samples as $\{x_j, p_j, y_j\}$, where lower case letters represent observations, indexed by subscripts. We use the well-known Nadaraya-Watson estimator:

$$E(Y|P) \approx \frac{\sum_{j=1}^{N} k_P(p_j - P) y_j}{\sum_{l=1}^{N} k_P(p_l - P)} = \sum_{j=1}^{N} \bar{k}_P(p_j - P) y_j, \tag{5}$$

where k_P is an appropriate kernel function and $\bar{k}_P(p_j - P) = \frac{k_P(p_j - P)}{\sum_{l=1}^{N} k_P(p_l - P)}$. Similarly, we can write:

$$E(E(Y|P)|X) \approx \sum_{k=1}^{N} \bar{k}_X(x_k - X) \sum_{j=1}^{N} \bar{k}_P(p_j - p_k) y_j, \tag{6}$$

where $\bar{k}_X(x_k - X) = \frac{k_X(x_k - X)}{\sum_{l=1}^{N} k_X(x_l - X)}$ and k_X is an appropriate kernel. Now, let's concatenate the observations into length N column vectors $\{\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{y}\}$ and define two $N \times N$ matrices K_X and K_P, the (j,k)'th entries of which are $\bar{k}_X(x_j - x_k)$ and $\bar{k}_P(p_j - p_k)$, respectively. Given the above, an estimate of κ is:

$$\kappa(Y|P|X) = \frac{\hat{V}(K_X K_P \boldsymbol{y})}{\hat{V}(\boldsymbol{y})}, \tag{7}$$

where $\hat{V}$ denotes the sample variance, defined as $\hat{V}(\boldsymbol{y}) = \frac{1}{N-1}\sum_{j=1}^{N}(y_j - \bar{y})^2$ with $\bar{y} = \frac{1}{N}\sum_{j=1}^{N} y_j$. Similarly, an estimate of γ is: $\gamma(P|X) = \frac{\hat{V}(K_X \boldsymbol{p})}{\hat{V}(\boldsymbol{p})}$.

In our implementation, we employed Gaussian kernel functions for k_X (and similarly k_P). I.e., $k_X(x_j - x_k) = \exp(-\frac{(x_j - x_k)^2}{h_X})$. Based on Silverman's rule of thumb we set the bandwidth as: $h_X = \hat{V}(\boldsymbol{x})/N^{0.2}$. Note that, this choice also ensures that the estimates are invariant to rescaling a variable.

3 Experimental Results

Data: We analyzed data from two public datasets, OASIS (oasis-brains.org) and ADNI (adni.loni.usc.edu), which contain brain MRI scans from healthy and demented subjects. We processed the T1-weighted structural brain MRIs using FreeSurfer (FS v5.1, surfer.nmr.mgh.harvard.edu) to obtain thickness measurements across the entire cortex, resampled onto a common template, *fsaverage*. FS also provides estimates of volumes for a range of cortical and sub-cortical structures, such as the hippocampus. The target variable we used was mini mental state exam (MMSE) score, which measures cognitive impairment and is associated with dementia, including Alzheimer's disease (AD). The OASIS sample consisted of young healthy subjects (YCN, N=200, 26.8 ± 9.7 years, 55% Female), old cognitively normal (OCN, clinical dementia rating, CDR, zero) subjects (N=135, 69.1 ± 13.8 y, 72% F) and AD patients (CDR>0, N=100, 76.8 ± 7.1 y, 59% F). We subdivided the OASIS OCN+AD sample (N=235) into five partitions (of equal size) for cross-validation. We call this the OASIS cross-validation sample. The entire ADNI sample contained N=810 (75.2 ± 6.9 y, 42% F) CN subjects, subjects with mild cognitive impairment and AD patients.

Machine Learning Algorithms: We explored two classes of publicly available algorithms to predict MMSE from brain MRI measurements. The first one is the Relevance Voxel Machine[1] [8] (RVoxM), which is an adaptation of a sparse Bayesian model, customized to handle image data. The second algorithm was the Support Vector Machine (SVM) with a radial basis function kernel[2]. We trained RVoxM and SVM to predict MMSE, based on FS-computed cortical thickness data. We also trained a separate SVM only on volumes of brain structures (saved as FS file *aseg.stats*), which we call SVM-aseg. We performed 5-fold cross-validation on the OASIS sample, where each of the five partitions was treated as the test sample in each fold, with the remaining subjects used for training. Thus, each OASIS subject was treated as a test case once, during which the ("out-of-bag") image-based prediction was computed. The Pearson correlation between out-of-bag predictions and ground truth values were 0.46,

[1] Downloaded from people.csail.mit.edu/msabuncu/sw/RVoxM/index.html
[2] Downloaded from csie.ntu.edu.tw/ cjlin/libsvm

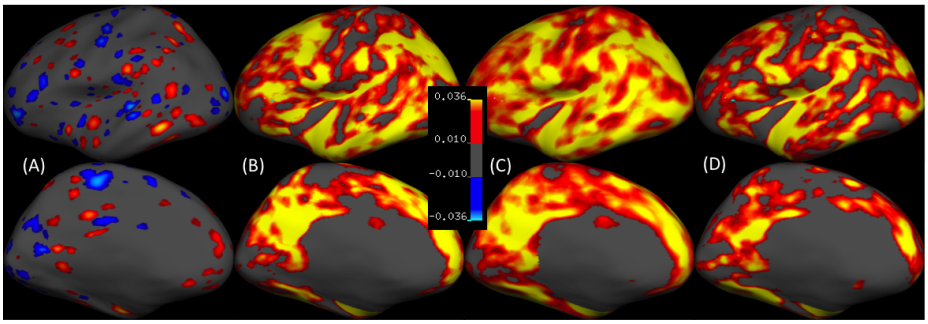


Fig. 1. All visualizations (in color) are on inflated *fsaverage* surface, a population average representation of the human cerebral cortex. Top and bottom rows show lateral and medial views, respectively. Only left hemispheres are shown. (A) Weights of RVoxM model trained on entire OASIS OCN+AD sample to predict MMSE from cortical thickness data. Note that most regions have no contribution to the model, i.e., have zero weight (shown in gray). (B) Captured correlation (κ) computed based on RVoxM's MMSE predictions on OASIS cross-validation sample. (C) κ-map for SVM's MMSE predictions on OASIS cross-validation sample (trained on cortical thickness). (D) κ-map for SVM-aseg's MMSE predictions on OASIS cross-validation sample.

0.53, and 0.35 (all P$< 10^{-10}$) for RVoxM, SVM and SVM-aseg, respectively. The ADNI data were only used for training to obtain prediction models.

Results: Fig. 1A visualizes the weights of the RVoxM model trained to predict MMSE from cortical thickness data on the OASIS sample. Because of RVoxM's sparsity assumption, most cortical regions have zero contribution to the model. We argue that this fact, along with the issues associated with interpreting the parameters of a discriminative model [11] makes it hard to make biological sense of this map and the SVM models. Moreover, we could not visualize the (non-linear) kernel SVM models, since there is no well-accepted strategy to do so. Fig. 1B-D illustrate maps of captured correlation (κ) computed with three different models trained on the OASIS cross-validation samples (based on out-of-bag predictions). All these maps bear a striking resemblance to AD-associated thinning maps reported in prior work [20]. Note MMSE is a variable strongly correlated with and used to clinically diagnose AD. The right-most map was actually computed based on an SVM model trained on the *aseg* features, which do not include regional cortical thickness measurements (although there is a variable that measures global cortical volume). There is strong agreement between these three κ-maps (pairwise Pearson correlations > 0.84, see Fig. 2D), which suggests that the captured correlation measure is robust to the variation in prediction algorithm and utilized image features.

Fig. 2A-C illustrate maps of GMC, ($\gamma(P|X)$, which ignores the ground truth variables Y) between cortical thickness values and the RVoxM predictions of MMSE. The correlation between the RVoxM-derived κ and γ-maps (Fig. 1-B and Fig. 2-A) is 0.97, providing evidence that the associations with the predicted values are informative about associations with the ground truth (thanks

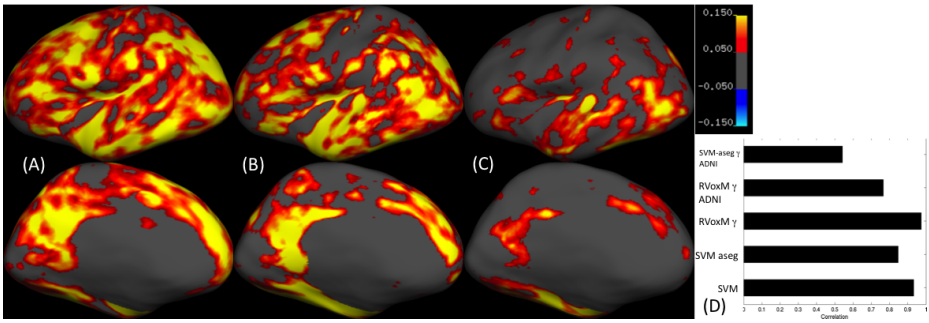


Fig. 2. (A) GMC (γ) between RVoxM's MMSE predictions and cortical thickness values computed on OASIS cross-validation sample. (B) γ-map between the ADNI RVoxM model's MMSE predictions and cortical thickness values computed on OASIS AD+CN sample. (C) γ-map between the ADNI RVoxM model's MMSE predictions and cortical thickness values computed on OASIS CN sample. (D) Pearson correlations of different maps with the RVoxM κ-map computed on OASIS cross-validation (shown in Fig. 1B). SVM, SVM-aseg, RVoxM γ, RVoxM γ ADNI, and SVM-aseg γ ADNI refer to maps of Fig. 1C, 1D, 2A, 2B, and 2C, respectively. For further details see caption of Fig.1.

to property **P6** of captured correlation). Fig. 2B-C were in fact computed using models trained on a separate dataset (ADNI). The map of Fig. 2C is particularly intriguing, as it was computed on healthy subjects (the OASIS young and old cognitively normal sample). Since this sample does not include subjects with dementia, there is little variation in the MMSE values (29.1 ± 1.1). However, the γ-map with the predicted MMSE scores demonstrate that, even in this healthy cohort, regions of potentially significant association with cognitive impairment can be detected. There is a correlation of 0.54 ($P < 1e-10$) between the map of Fig. 2C and the benchmark map of Fig. 1B. This result demonstrates the robustness of the proposed measure with respect to substantial variation in the data, since both the training and testing data are different between the analyses.

4 Conclusion

We proposed a novel measure, called captured correlation, to quantify associations between input features and the target variable, as captured by the prediction model. We applied this measure to image-based prediction models and demonstrated that captured correlation yields biologically meaningful maps that are robust to the choice of learning algorithm. We showed that under certain assumptions, captured correlation is proportional to the association between features and predictions. Intriguingly, this perspective provides a theoretical justification for examining associations with predictions, in the absence of ground truth labels. For example, one can analyze large, unlabeled datasets in order to identify potentially relevant areas, which could then be further interrogated on labeled datasets. Our approach can also be used to examine and prioritize multivariate relationships, such as the association between multiple image features and the target variable. Future work will pursue these interesting directions.

References

1. Breiman, L.: Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science 16(3), 199–231 (2001)
2. Kawasaki, Y., et al.: Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. Neuroimage 34(1) (2007)
3. Davatzikos, C., et al.: Individual patient diagnosis of ad and ftd via high-dimensional pattern classification of mri. Neuroimage 41(4), 1220–1227 (2008)
4. Klöppel, S., et al.: Automatic detection of preclinical neurodegeneration presymptomatic huntington disease. Neurology 72(5), 426–431 (2009)
5. Zhang, D., et al.: Multimodal classification of alzheimer's disease and mild cognitive impairment. Neuroimage 55(3), 856–867 (2011)
6. Plant, C., et al.: Automated detection of brain atrophy patterns based on mri for the prediction of alzheimer's disease. Neuroimage 50(1), 162–174 (2010)
7. Ashburner, J., Friston, K.J.: Voxel-based morphometry–the methods. Neuroimage 11(6), 805–821 (2000)
8. Sabuncu, M.R., Van Leemput, K.: The Relevance Voxel Machine (RVoxM): A self-tuning bayesian model for informative image-based prediction. IEEE Transactions on Medical Imaging (2012)
9. Gaonkar, B., Davatzikos, C.: Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. NeuroImage 78, 270–283 (2013)
10. Konukoglu, E., Ganz, M., Van Leemput, K., Sabuncu, M.R.: On feature relevance in image-based prediction models: An empirical study. In: Wu, G., Zhang, D., Shen, D., Yan, P., Suzuki, K., Wang, F. (eds.) MLMI 2013. LNCS, vol. 8184, pp. 171–178. Springer, Heidelberg (2013)
11. Haufe, S., et al.: On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage 87, 96–110 (2014)
12. Golland, P.: Discriminative direction for kernel classifiers. Advances in Neural Information Processing Systems 1, 745–752 (2002)
13. Strobl, C., et al.: Conditional variable importance for random forests. BMC Bioinformatics 9(1), 307 (2008)
14. Zien, A., Krämer, N., Sonnenburg, S., Rätsch, G.: The feature importance ranking measure. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009, Part II. LNCS (LNAI), vol. 5782, pp. 694–709. Springer, Heidelberg (2009)
15. Meinshausen, N., Bühlmann, P.: Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72(4), 417–473 (2010)
16. Goldstein, A., et al.: Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. J. of Comp. and Graph. Stat. (2014)
17. Dosenbach, N.U.F., et al.: Prediction of individual brain maturity using fMRI. Science 329(5997), 1358 (2010)
18. Sonnenburg, S., et al.: Poims: positional oligomer importance matricesunderstanding support vector machine-based signal detectors. Bioinformatics 24(13), i6–i14 (2008)
19. Zheng, S., et al.: Generalized measures of correlation for asymmetry, nonlinearity, and beyond. J. of the American Statistical Association 107(499), 1239–1252 (2012)
20. Dickerson, B., et al.: The cortical signature of alzheimer's disease. Cerebral Cortex 19(3), 497–510 (2009)

3D Spine Reconstruction of Postoperative Patients from Multi-level Manifold Ensembles

Samuel Kadoury[1,2], Hubert Labelle[2], and Stefan Parent[2]

[1] MEDICAL, Polytechnique Montreal, Montreal, QC, Canada
samuel.kadoury@polymtl.ca
[2] Sainte-Justine Hospital Research Center, Montreal, QC, Canada

Abstract. The quantitative assessment of surgical outcomes using personalized anatomical models is an essential task for the treatment of spinal deformities such as adolescent idiopathic scoliosis. However an accurate 3D reconstruction of the spine from postoperative X-ray images remains challenging due to presence of instrumentation (metallic rods and screws) occluding vertebrae on the spine. In this paper, we formulate the reconstruction problem as an optimization over a manifold of articulated spine shapes learned from pathological training data. The manifold itself is represented using a novel data structure, a multi-level manifold ensemble, which contains links between nodes in a single hierarchical structure, as well as links between different hierarchies, representing overlapping partitions. We show that this data structure allows both efficient localization and navigation on the manifold, for on-the-fly building of local nonlinear models (manifold charting). Our reconstruction framework was tested on pre- and postoperative X-ray datasets from patients who underwent spinal surgery. Compared to manual ground-truth, our method achieves a 3D reconstruction accuracy of 2.37 ± 0.85mm for postoperative spine models and can deal with severe cases of scoliosis.

1 Introduction

Spinal deformity pathologies such as adolescent idiopathic scoliosis are three-dimensional (3D) deformations of the trunk, described as a lateral deviation of the spine combined with asymmetric deformation of the vertebrae. Modalities such as MRI or CT are limited for pre- and postoperative assessment since they require the patient to be lying in a prone position during acquisition. For this reason, biplanar radiography is still the imaging technique which is most frequently used for the 3D clinical assessment of spinal deformities, as it allows acquiring the image in a natural standing posture with very little radiation.

Several attempts were made to reconstruct the spine from biplanar X-rays. *A priori* knowledge of the vertebral shapes was used with morphologic descriptors to estimate the geometrical spine model and refined with projected silhouettes [1]. This approach was improved by adding inference-based adjustments to obtain an accurate estimate of the vertebra's orientation and 3D location [2]. Humbert et al. [3] proposed to evaluate a parametric model based on the spinal centerline,

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 361–368, 2014.

Moura et al. [4] inferred an articulated model of the spine based on splines, and Boisvert et al. [5] formulated the estimation of the spine shape as a second-order cone program. Kadoury et al. [6] introduced a statistical and image-based approach to reconstruct the spine using geometrical models. However, all of these methods were designed for the reconstruction from preoperative X-rays.

Surgical treatment involves correcting the scoliotic curves with pre-shaped metal rods anchored on the vertebrae with screws. Few methods have focused on the 3D reconstruction from instrumented spines, even though postoperative 3D analysis is crucial to assess a treatment's efficacy. In [7], a multilevel statistical model was proposed to reconstruct the spine in postoperative patients. Still, the method remains highly supervised and assumes a linear statistical distribution of the underlying pathological variations. In contrast, manifold learning is based on the premise that data are often of artificially high dimension and can be embedded in a lower dimensional space. However most global approaches (LLE, ISOMAP) fail to adequately model closed distributions by unwrapping the manifold, thus changing the intrinsic topology. On the other hand, graph-search approaches preserve the manifold, avoiding losing continuity during the embedding [8]. This could add robustness to the inference process of unseen cases from a low-dimensional manifold embedded in the ambient space.

We propose a novel biplanar 3D reconstruction method of the instrumented spine, where the shape reconstruction task is formulated as optimizing an energy function over the manifold of uninstrumented spine shapes. The energy function is designed such that it is robust to instrumentation, leading to solutions which fit inside the domain of valid anatomical configurations. The manifold is learned from a training set of visual hull reconstructions from spine X-ray images. To this manifold we attach a vector field of generating parameters for articulated poses and shapes. Our contributions are two-fold. First, we propose an approach to build local charts whenever it is required on the manifold, approximating the tangent space around a point and maximizing the accuracy of the nonlinear approximation. Our approach combines the construction of a neighbourhood graph and learning the manifold onto this graph, to effectively infer new models. Second, we propose a spine reconstruction method from X-ray images, where the anatomy of the spine is occluded by instrumentation (screws and rods).

2 Method

Given a set of calibrated biplanar X-rays, a visual hull reconstruction of the global spine shape is obtained from the silhouettes of the anterior portion of the spine extracted from both images (frontal and lateral) using a Hessian filtering technique [9]. Our aim is to create a multi-level manifold ensemble from a training set of visual hulls (Fig. 1), so to efficiently navigate towards the best match during the spine reconstruction process for an unseen set of X-rays.

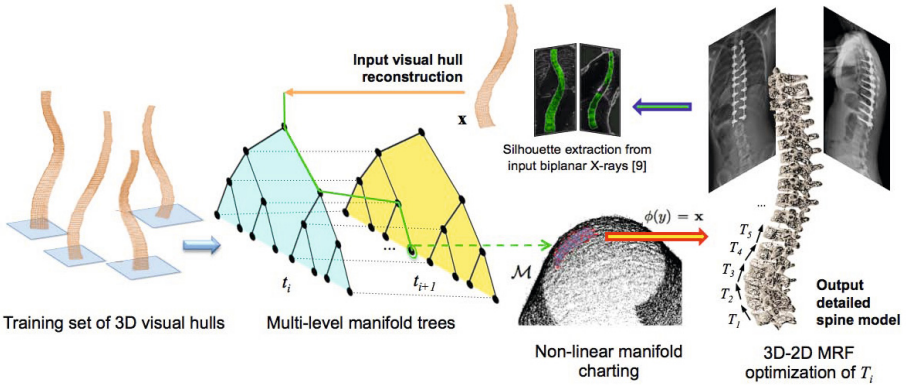


Fig. 1. Illustration of the multi-level manifold ensembles for the 3D reconstruction of postoperative spines. The data structure allows both vertical and horizontal displacements between tree nodes given an input visual hull reconstruction of the spine.

2.1 Building the Multi-level Manifold Ensemble

The first step is to learn an ensemble of randomized space partioning trees that are connected between each other. Essentially, the trees can be viewed as defining an adaptive grid on the ambient space, in a manner similar to k-d trees. The ensemble is a set, $\mathcal{T}$, of binary trees $t_i \in \mathcal{T}$ which hierarchically partitions the ambient data space $\mathbb{R}^D$. We train each tree with the same dataset $\mathcal{X} = \{\mathbf{x}_i\}$, $\mathbf{x}_i \in \mathbb{R}^D$ of global spine visual hulls of dimension D. In our application, we assume the samples $\mathbf{x}_i$ to lie on a d-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^D$ with $d < D$. The parameters $\Theta_j = (\theta_j, \tau_j)$ for each node j define a separating hyperplane in the ambient space $\mathbb{R}^D$ by its unit normal vector $\theta_j \in \mathbb{R}^D$ and a threshold $\tau_j \in \mathbb{R}$, set as the average of the two projected samples with the largest distance on the manifold. The data assigned to each node, $\mathcal{X}_j$, is partitioned into two subsets: $\mathcal{X}_j^L$ and $\mathcal{X}_j^R$, depending on the value of the split function $h(\mathbf{x}, \Theta_j) \in \{0, 1\}$. The split functions take the form:

$$h(\mathbf{x}, \Theta_j) = \mathrm{I}(\mathbf{x}^T \theta_j > \tau_j) \tag{1}$$

where I is the indicator function that penalizes points further away from the manifolds. The set $\mathcal{X}_j^L$ contains samples $\mathbf{x} \in \mathcal{X}$ for which $h(\mathbf{x}, \Theta_j) = 0$, the set $\mathcal{X}_j^R$ contains those for which $h(\mathbf{x}, \Theta_j) = 1$. To find Θ_j, we sample a random subset, $\mathcal{D}_j \subset \mathcal{X}_j$, sample a point $\mathbf{x}_k \in \mathcal{D}_j$ and find the most distant point in $\mathcal{D}_j$:

$$\mathbf{x}_l = \underset{\mathbf{x} \in \mathcal{D}_j}{\operatorname{argmax}} \|\mathbf{x}_l - \mathbf{x}_k\|. \tag{2}$$

The normal θ_j to our hyperplane is the unit length vector between these two points: $\theta_j = (\mathbf{x}_l - \mathbf{x}_k)/(\|\mathbf{x}_l - \mathbf{x}_k\|)$.

The second step learns the graphs, composed of a set of tree nodes, $\mathcal{V}$, and a set of directed edges, $\mathcal{E}$. Here, we denote nodes as $v_i^t \in \mathcal{V}$ with t the tree, and

i the node index respectively, while an edge set is composed of all parent-child tree edges $\mathcal{E}_t$ along with edges between trees $\mathcal{E}_{s,t}$. A set of edges is defined as:

$$\mathcal{E} = (\bigcup_{t \in \mathcal{T}} \mathcal{E}_t) \cup (\bigcup_{(s,t) \in \mathcal{T}} \mathcal{E}_{s,t}). \tag{3}$$

The idea here is that two nodes v_i^s and v_j^t in trees s and t are connected if the regions they define intersect. Exact computation of these intersections is too expensive in high dimensions, even in the case of linear splits. Instead, we use the data samples to estimate intersections, and connect nodes v_i^s and v_j^t if the intersection of their sample sets $\mathcal{D}_i^s$ and $\mathcal{D}_j^s$ is non-empty. The weight $w_{(v_i^s, v_j^t)}$ of each directed edge is set proportional to $\mathcal{D}_i^s \cap \mathcal{D}_j^s$.

2.2 Optimization on the Manifold

Once the multi-level manifolds are trained, our aim, given a test case, is to optimize an energy function f defined on points that lie on a manifold $\mathcal{M}$. We first find an initial solution by traversing the ensemble, both horizontally and vertically (within and between trees), using an optimized k-d search with aligned trees. Upon reaching a leaf node, we query the local neighborhood with the horizontal connections and build a nonlinear chart on $\mathcal{M}$.

Generating an Initial Solution. In order to minimize an energy function f with the multi-level manifold structure, the first step consists of generating a set of initial points on the manifold. These potential nodes will provide a pool of candidates which will be close to the global solution. Our aim is not to obtain a single point, but rather a group of points on the manifold used later to chart a non-linear map. To overcome the overall complexity of the multi-level trees and facilitate searches to find the nearest neighbour to a query, we employ a strategy where prior to searching the trees, binary trees $t_i \in \mathcal{T}$ are aligned by their principal components via a PCA approach. Data is therefore split up in the tree by hyperplanes perpendicular to the principal axes. To achieve this, the datasets are first translated so that their centroids coincide at the origin. Then, we construct the matrix $U = \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$, with N the size of t_i. The eigenvectors of U are the principal axes of the dataset, and the eigenvalues are referred to as the principal moments which are used to rotate the trees $t_i \in \mathcal{T}$.

Given a query vector, a descent down the trees leads to a single leaf node. The data point associated with the node v_j^t is the first candidate for the nearest neighbour. However this point will not necessarily be the nearest neighbour to the query vector; it must be followed by a process of horizontal and vertical moves, in which other cells are searched for better candidates. The recommended method is priority search [10] in which the cells are searched in the order of their distance from the query point. This may be accomplished efficiently using a priority tree for ordering the cells, in which the cells are numbered in the order of their distance from the query vector. The search terminates when there are no more cells within the distance defined by the best point found so far.

At the end of the process, the minimizing cost function yields a set of leaf averages $\bar{\mathbf{x}}_j^t$ which are used as points to create a non-linear chart on the manifold $\mathcal{M}$. Hence, this feature represents a clear advantage to methods using graph-building approaches, since charts are computed on the fly over the training data.

Nonlinear Charting. All the leaf averages $\bar{\mathbf{x}}_j^t$ define a neighbourhood in the local nonlinear charting step of the manifold $\mathcal{M}$, finding a mapping function around a query. The seeds are expanded to their neighborhood using a random walk of horizontal moves. The walk carries on until a given number of nodes W has been reached. The parameter W controls the local chart size and must therefore be chosen carefully, depending on a given optimization problem. We propose to use the nodes reached by the walk as samples for training $\phi(y)$, which provides the transformation from a nonlinear chart near $\mathbf{x}$ on the manifold $\mathcal{M}$ at $y \in \mathbb{R}^d$ to the ambient space, such that $\phi(y) = \mathbf{x}$ where y contains the non-linear dimensionally components. We use the basis of the tangent plane to estimate $\phi(y)$ in d-dimensional space, where in differential geometry, the tangent space is used to compute a local chart. The mapping of $\phi(y) = \mathbf{x}$ estimates the relationship between the D-space and manifold $\mathcal{M}$ as a joint distribution. The non-linear chart should follow a conditional expectation which captures the trend within a local neighborhood of the manifold. Gaussian kernels G estimate densities in the conditional expectation setting [11]:

$$\phi(y_i) = \underset{\mathbf{x}_i}{\operatorname{argmin}} \frac{\sum_{j \in \mathcal{N}(i)} G(y_i, y_j) \|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sum_{j \in \mathcal{N}(i)} G(y_i, y_j)} \tag{4}$$

which integrates the distance $\|\mathbf{x}_i - \mathbf{x}_j\|^2$, that: 1) acts as the similarity metric, 2) avoids giving more weight to the neighbors that are further away within a the neighbourhood $\mathcal{N}(i)$ and, 3) updates $\phi(y_i)$ using the closest neighbors of point y_i in the manifold space. This constrains the regression to be valid for all the neighbourhood of the manifold chart around y_i as it preserves locality in $\mathbf{x}_i$. The chart provides a non-linear parameterization of the space, where outside this range, a new chart is recomputed around the new initial solution.

2.3 3D Reconstruction of Postoperative Patients

We formulate the 3D reconstruction process of postoperative spines, where the anatomy is partly obscured with rods and screws on the X-ray images, by an optimization over the manifold of visual hulls from uninstrumented spines. The energy term we seek to optimize is asymmetric, meaning that the solution that is obtained from the manifold must lie inside the visual hull of the instrumented input. Our model is learned from visual hull shapes (vectorized as $\mathbf{x}$) which are smoothed and subsampled. The function f used to navigate throughout the manifold and find the closest points to the query $\mathbf{x}$ is formulated as:

$$f(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{|\mathbf{x}|} k \circ |(x_i - y_i)(1 + \mathrm{I}(x_i < y_i)\beta)| \tag{5}$$

which defines the dissimilarity between input visual hull $\mathbf{x}$ and points $\mathbf{y}$ on the manifold. Here, $k(\cdot)$ is a kernel function reducing the influence of outliers, and $\mathrm{I}(\cdot)$ is the indicator function. The function f induces a penalty of β for points falling outside the input reconstruction $\mathbf{x}$ i.e. the uninstrumented visual hull must lie within the input shape. We compute initial solutions on the manifold, followed by nonlinear charting using Eq. (4) over a neighbourhood from which a geometrical spine model is built.

Once the closest node on the manifold is identified, we revert to the ambient space for the final refinements brought to the detailed spine model, by adjusting the parametric spine model to its projection on the biplanar X-rays. We build the parametric body model from the vector field of generating parameters attached to the manifold neighborhood found previously. This models the spine as an array of local intervertebral rigid transformations $A = [T_1, T_2, \ldots, T_m]$, with m the number of intervertebral transformations and $T_i = \{R, t\}$ a rigid transform. We propose to use a Markov Random Field (MRF) to refine A, which was obtained from the manifold. In our formulation, the energy includes data term $\mathbf{g} = \{g_i(\cdot)\}$ (unary potentials) associated to each vertex of the MRF and a regularization term $\mathbf{f} = \{f_{ij}(\cdot,\cdot)\}$ (pairwise potentials) associated to the edges. The first encodes the geodesic active contours of the projected 3D vertebral meshes with the X-ray gradient images, whereas the later ones act as a regularizer by restraining large transformation differences between two vertebral levels (e.g. T_4 and T_5), introducing coherency in the minimization of the discrete MRF:

$$MRF(\mathbf{g}, \mathbf{f}) = \min \sum_{i=1}^{m} g_i(\mathbf{u}_i) + \sum_{i=1}^{m} \sum_{j \in \mathcal{N}(i)} f_{ij}(\mathbf{u}_i, \mathbf{u}_j). \tag{6}$$

Here, $\mathbf{u}_i, \mathbf{u}_j \in L$ are the labels assigned to transformations $T_i, T_j \in A$ respectively, with transformation T_i being moved by assigning them different labels $\mathbf{u}_i \in L$ (L is the label space) until an optimal configuration is found. To reach such an optimal configuration, we define an energy term that will be minimized using an optimization algorithm. We adopt FastPD (Primal-Dual) to optimize the aforementioned non-submodular MRF.

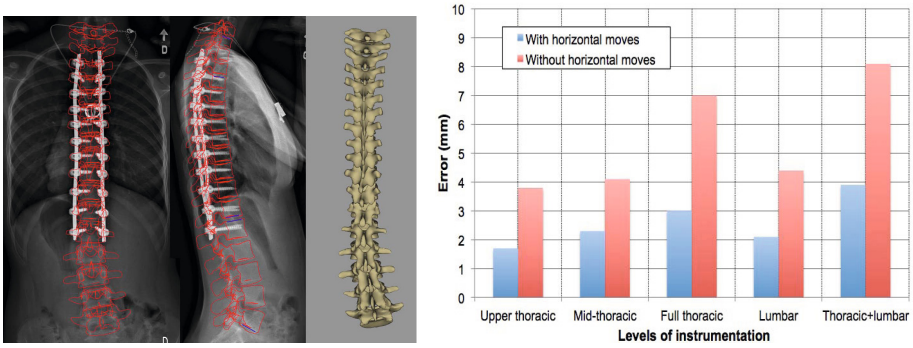
3 Experiments

We tested the multi-level manifold structure for the 3D reconstruction of spine models with instrumentation, by handling occlusions from X-rays. The manifold was built from 843 scoliotic spines demonstrating several types of deformities. For each spine in the dataset, both a visual hull reconstruction and a high-resolution geometrical model was obtained from calibrated biplanar X-rays. The models include 12 thoracic and 5 lumbar vertebrae, each represented by 6 landmarks (4 pedicle extremities and 2 endplate center points), yielding a total 102 landmarks per model. An atlas of 17 generic prior vertebra mesh models obtained from serial CT-scans were fitted to the landmark-based models using FFD.

We first evaluate the performance on preoperative patients with scoliosis. Ten pairs of preoperative biplanar X-rays taken from patients with mild deformities

Table 1. Reconstruction results of pre- and postoperative spines. Results of the proposed multi-level manifold (MLM) are compared to standard PCA and a recently proposed approach using locally linear embeddings (LLE) [6].

	Preoperative X-rays ($n = 10$)						Postoperative X-rays ($n = 20$)					
dim.	2D RMS difference			3D RMS difference			2D RMS difference			3D RMS difference		
	PCA	LLE	**MLM**	PCA	LLE	**MLM**	PCA	LLE	**MLM**	PCA	LLE	**MLM**
1	4.77	2.72	**2.05**	5.04	2.97	**2.39**	6.37	3.44	**2.38**	6.72	3.86	**2.59**
5	4.18	2.53	**1.98**	4.52	2.85	**2.22**	5.74	3.31	**2.21**	6.16	3.80	**2.46**
10	3.71	2.07	**1.90**	4.10	2.33	**2.09**	5.53	3.18	**1.96**	5.95	3.68	**2.34**
15	3.64	2.02	**1.93**	4.04	2.28	**2.15**	5.52	3.14	**2.04**	5.87	3.63	**2.39**
20	3.59	2.10	**1.95**	4.01	2.39	**2.20**	5.47	3.22	**2.07**	5.85	3.72	**2.41**



(a) (b)

Fig. 2. (a) Sample 3D reconstruction result from postoperative frontal and lateral X-rays, yielding high-resolution geometrical models. (b) Errors with 5 different instrumentation configurations, comparing optimization with and without horizontal moves.

(Cobb angle 15 - 30°) were used to quantify the 3D accuracy. This enables to assess the errors under normal settings, without any occlusion from instrumentation. For each case, differences between the proposed method and landmarks identified manually by a radiology expert were computed. Table 1 presents these results, demonstrating performances similar to state-of-the-art [6].

We then evaluated the method on twenty ($n = 20$) pairs of instrumented biplanar X-rays, where 3D spine reconstructions were generated by learning a manifold of visual hulls rendered from uninstrumented spines. Sample results are shown in Fig. 2(a). In Table 1, we compare results to both a PCA and a single spine manifold approach [6], prior to MRF optimization. Different values of d were tested in order to evaluate the effect of the manifold dimensionality. With $d = 10$, the mean difference in 3D landmark locations for all 20 patients was of 2.57 ± 0.73mm for thoracic vertebrae and 2.12 ± 0.67mm for lumbar vertebrae. To evaluate robustness with respect to different instrumentation strategies, we measured the reconstruction accuracy over five different types of surgical paradigms, based on instrumented levels (thoracic and lumbar). Fig. 2(b) shows these results, which also show the benefit of integrating horizontal moves in the ensemble search when different anatomical regions are occluded by metal rods.

4 Conclusion

In this paper, we proposed an unsupervised method to perform the 3D reconstruction of a spine geometry when surgical instrumentation is visible on biplanar X-rays. Our approach is based on multi-level manifold ensembles, which enable an efficient navigation on a low-dimensional domain to infer the closest match to a training set. Results show that this model allows reconstruction accuracies similar to gold-standard. Because the 3D reconstructions are obtained without any user supervision, the proposed approach could be transposed to clinical practice and used in the context of multi-centre evaluations of surgical practices.

References

1. Pomero, V., Mitton, D., Laporte, S., de Guise, J.A., Skalli, W.: Fast accurate stereoradiographic 3D-reconstruction of the spine using a combined geometric and statistic model. Clinical Biomech. 19, 240–247 (2004)
2. Dumas, R., Blanchard, B., Carlier, R., de Loubresse, C.G., Huec, J.C.L., Marty, C., Moinard, M., Vital, J.M.: A semi-automated method using interpolation and optimisation for the 3D reconstruction of the spine from bi-planar radiography: a precision and accuracy study. Med. Biol. Eng. Comput. 46, 85–92 (2008)
3. Humbert, L., de Guise, J., Aubert, B., Godbout, B., Skalli, W.: 3D reconstruction of the spine from biplanar X-rays using parametric models based on transversal and longitudinal inferences. Med. Eng. Phy. 31(6), 681–687 (2009)
4. Moura, D., Boisvert, J., Barbosa, J., Labelle, H., Tavares, J.: Fast 3D reconstruction of the spine from biplanar radiographs using a deformable articulated model. Med. Eng. Phy. 33, 924–933 (2011)
5. Boisvert, J., Moura, D.: Interactive 3D reconstruction of the spine from radiographs using a statistical shape model and second-order cone programming. In: 33rd Annual International Conference of the IEEE EMBS, pp. 5726–5729 (2011)
6. Kadoury, S., Cheriet, F., Labelle, H.: Personalized X-ray 3D reconstruction of the scoliotic spine from statistical and image models. IEEE Trans. Med. Imag. 28, 1422–1435 (2009)
7. Lecron, F., Boisvert, J., Mahmoudi, S., Labelle, H., Benjelloun, M.: Fast 3D spine reconstruction of postoperative patients using a multilevel statistical model. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part II. LNCS, vol. 7511, pp. 446–453. Springer, Heidelberg (2012)
8. Pitelis, N., Russell, C., Agapito, L.: Learning a manifold as an atlas. In: CVPR, pp. 1–8 (2013)
9. Kadoury, S., Cheriet, F., Labelle, H.: Segmentation of scoliotic spine silhouettes from enhanced biplanar X-rays using a prior knowledge Bayesian framework. In: Proc. ISBI, pp. 478–481 (2009)
10. Silpa-Anan, C., Hartley, R.: Optimised KD-trees for fast image descriptor matching. In: CVPR, pp. 1–8 (2008)
11. Davis, B., Fletcher, P., Bullitt, E., Joshi, S.: Population shape regression from random design data. In: ICCV, pp. 1–7 (2007)

Scalable Histopathological Image Analysis via Active Learning

Yan Zhu[1], Shaoting Zhang[2], Wei Liu[3], and Dimitris N. Metaxas[1]

[1] Department of Computer Science, Rutgers University, Piscataway, NJ, USA
[2] Department of Computer Science, University of North Carolina at Charlotte, NC, USA
[3] IBM T.J. Watson Research Center, NY, USA

Abstract. Training an effective and scalable system for medical image analysis usually requires a large amount of labeled data, which incurs a tremendous annotation burden for pathologists. Recent progress in active learning can alleviate this issue, leading to a great reduction on the labeling cost without sacrificing the predicting accuracy too much. However, most existing active learning methods disregard the "structured information" that may exist in medical images (*e.g.*, data from individual patients), and make a simplifying assumption that unlabeled data is independently and identically distributed. Both may not be suitable for real-world medical images. In this paper, we propose a novel batch-mode active learning method which explores and leverages such structured information in annotations of medical images to enforce diversity among the selected data, therefore maximizing the information gain. We formulate the active learning problem as an adaptive submodular function maximization problem subject to a partition matroid constraint, and further present an efficient greedy algorithm to achieve a good solution with a theoretically proven bound. We demonstrate the efficacy of our algorithm on thousands of histopathological images of breast microscopic tissues.

1 Introduction

Recent development of microscopical acquisition technology enables computerized analysis of histopathological images [9]. For example, in the context of breast cancer diagnosis, plenty of systems have been designed to conduct automatic and accurate analysis of high-resolution images digitized from tissue histopathology slides, where well-known machine learning and image processing techniques [12,3,4] have been exploited. Particularly, supervised learning models such as Support Vector Machines (SVMs) [13] have been extensively employed, because they are able to effectively bridge the so-called "semantic gap" between histopathological images and their diagnosis information [3,6,9]. To train an accurate prediction model under a supervised manner, it is usually necessary to require a large amount of labeled data, *e.g.*, manual annotations from domain experts or pathologists. However, acquiring sufficient high-quality annotations is a very expensive and tedious process. To alleviate this issue and reduce the labeling cost, active learning [14] has been suggested to intelligently select a small yet informative subset of the whole database, which requires only a few labeling operations from domain experts to build an accurate enough prediction model yet with a low training cost.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 369–376, 2014.

Active learning has been widely investigated in the machine learning community, aiming for progress in both theoretical aspects, *e.g.*, sample complexity bounds [1], and approaching practical applications, *e.g.*, image [10] and text [15] classification and retrieval (the related work in active learning is briefly described below). However, for histopathological images, previous active learning methods have two main shortcomings: 1) Almost all of them assume that unlabeled data samples are *independently and identically distributed* (I.I.D.), which is not necessarily suitable for histopathological images. In fact, for each patient there are usually several images available which share common pathological characteristics, *e.g.*, images from different ROIs. Obviously, there are considerable correlations among such image samples. 2) Even if the I.I.D. property holds, previous active learning methods may disregard the structured information of histopathological images, *e.g.*, patient identity, which is easy to obtain but could be crucial for active learning to enforce diversity during sample selection.

In this work, we propose a novel batch mode active learning approach which is specifically designed for histopathological image analysis by leveraging structured information to enforce diversity during intelligent sample selection. We formulate the active learning problem (essentially the sample selection problem) as a constrained submodular optimization problem and present a greedy algorithm to efficiently solve it. Notably, we provide a theoretical bound characterizing the quality of the submodular active learning strategy, which guarantees that our proposed greedy algorithm approximates the optimal batch mode active learning strategy for the adaptive submodular function maximization problem with a partition matroid constraint. In practice, our active learning driven histopathological image analysis approach outperforms state-of-the-art methods to tackle histopathological image analysis. We perform experiments on a large database of histopathological images with high-dimensional features. The experimental results demonstrate the efficacy of our approach, which achieves 83% prediction accuracy with merely 100 labeled samples among more than two thousand images (*i.e.*, less than 5% training data). This accuracy is 11% higher than passive learning and 6% higher than state-of-the-art active learning methods.

Related Work in Active Learning. Active learning can be considered as a combinatorial optimization problem which is typically difficult to exactly solve, so a variety of heuristics have been resorted to. For example, a number of active learning algorithms relax the original combinatorial problem involving discrete constraints to a continuous optimization problem, and then employ regular convex or non-convex optimization techniques to solve the relaxed problem. These algorithms usually suffer from prohibitively high computational complexities, and the deviation from the solution of the relaxed problem to that of the original problem remains unknown. In contrast, some latest work casts active learning problem into a submodular set function maximization problem which is direct combinatorial optimization. While maximizing a submodular function appears NP-hard, a landmark result from Nemhauser *et al.* [5] certifies that a simple greedy optimization scheme is able to achieve the $(1 - \frac{1}{e})$-approximation for the cardinality constraint and the $(\frac{1}{p+1})$-approximation for p matroid constraints, respectively. Built on this theoretic finding, Chen and Krause [2] propose a nearly optimal batch mode active learning strategy by applying an adaptive submodular optimization scheme [8]. Motivated by this line of submodular optimization techniques, our active

learning method firstly explores and leverages structured information of histopathological images through imposing a partition matroid constraint on active learning.

2 Approach

2.1 Problem Definition

Given an unlabeled dataset $\mathcal{U} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$, each data sample $\boldsymbol{x}_i \in \mathcal{U}$ carries a random label variable $y_i \in \mathcal{Y}$ ($\mathcal{Y} = \{1, -1\}$) in our binary classification task for which the positive label '1' implies 'benign' and the negative label '-1' implies 'actionable'. Assume that there exists a joint probability distribution $P(\boldsymbol{y}_\mathcal{U})$ of the labels of the samples in $\mathcal{U}$, where $\boldsymbol{y}_\mathcal{U} = [y_1, \cdots, y_n]^\top \in \mathcal{Y}^n$. Batch mode active learning selects a small subset of $\mathcal{U}$, queries their labels from experts, and then trains a classifier using the chosen labeled samples. To be specific to histopathological image analysis, batch mode active learning works as follows: whenever a batch of k unlabeled images $\mathcal{B} \subseteq \mathcal{U}$ ($|\mathcal{B}| = k$) are selected, their associated labels $\boldsymbol{y}_\mathcal{B} \in \mathcal{Y}^k$ are requested from the diagnosis of pathologists and acquired simultaneously; the obtained labels are used to select next batches of images iteratively until the needed classification (*i.e.*, predicting 'benign' or 'actionable') accuracy is achieved.

2.2 Adaptive Submodular Optimization

Our goal is to learn a classifier $h : \mathcal{U} \to \mathcal{Y}$ from a set $\mathcal{H}$ of finite hypotheses. We write $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\} \subseteq \mathcal{U} \times \mathcal{Y}$ to denote the set of observed sample-label pairs. We define $\mathcal{H}(\mathcal{S}) = \{h \in \mathcal{H} : y_i \equiv h(\boldsymbol{x}_i), \forall (\boldsymbol{x}_i, y_i) \in \mathcal{S}\}$ to denote the reduced hypothesis space consistent with the observed sample-label pairs in $\mathcal{S}$. We then define and aim to maximize the objective set function $f : 2^{\mathcal{U}\times\mathcal{Y}} \to \mathbb{R}$ as

$$f(\mathcal{S}) = |\mathcal{H}| - |\mathcal{H}(\mathcal{S})|, \tag{1}$$

where the operator $|\cdot|$ outputs the cardinality of an input set. In this paper, we study hyperplane hypotheses in the form of $h(\boldsymbol{x}) = \mathrm{sgn}(\boldsymbol{w}^\top \boldsymbol{x})$ in which the sign function $\mathrm{sgn}(x)$ returns 1 if $x > 0$ and -1 otherwise. Intuitively, the function $f(\mathcal{S})$ measures the number of hypotheses eliminated by the observed labeled data in $\mathcal{S}$. As a matter of fact, f satisfies the following properties:

- $f(\varnothing) = 0$; (**Normalized**)
- for any $\mathcal{S}_1 \subseteq \mathcal{S}_2 \subseteq \mathcal{U} \times \mathcal{Y}$, $f(\mathcal{S}_1) \leq f(\mathcal{S}_2)$; (**Monotonic**)
- for any $\mathcal{S}_1 \subseteq \mathcal{S}_2 \subseteq \mathcal{U} \times \mathcal{Y}$ and $(\boldsymbol{x}, y) \in (\mathcal{U} \times \mathcal{Y}) \backslash \mathcal{S}_2$, we have $f(\mathcal{S}_2 \cup \{(\boldsymbol{x}, y)\}) - f(\mathcal{S}_2) \leq f(\mathcal{S}_1 \cup \{(\boldsymbol{x}, y)\}) - f(\mathcal{S}_1)$; (**Submodular**)
- for an unlabeled sample $\boldsymbol{x}$ and an observed data subset $\mathcal{S} \subseteq \mathcal{U} \times \mathcal{Y}$, define the conditional expected marginal gain of $\boldsymbol{x}$ with regard to $\mathcal{S}$ as

$$\Delta_f(\boldsymbol{x} \mid \mathcal{S}) = \sum_{y \in \mathcal{Y}} P(y_i = y \mid \mathcal{S})[f(\mathcal{S} \cup \{(\boldsymbol{x}, y)\}) - f(\mathcal{S})], \tag{2}$$

  and then the function f along with the distribution $P(\boldsymbol{y}_\mathcal{U})$ is called adaptive submodular if $\Delta_f(\boldsymbol{x} \mid \mathcal{S}_2) \leq \Delta_f(\boldsymbol{x} \mid \mathcal{S}_1)$ holds for any $\mathcal{S}_1 \subseteq \mathcal{S}_2 \subseteq \mathcal{U} \times \mathcal{Y}$ and $P(\mathcal{S}_2) > 0$. (**Adaptive Submodular** [8])

To work under the batch mode setting, the *BatchGreedy* algorithm [2] generalizes the conditional marginal benefit in Eq. (2) to allow for conditioning on a set of selected but not yet observed sample-label pairs within the current batch. *BatchGreedy* greedily selects the samples within each batch and assembles batches in a sequential manner. Specifically, *BatchGreedy* selects the i-th sample in the j-th batch as follows:

$$\boldsymbol{x}^* = arg\max_{\boldsymbol{x}\in\mathcal{U}} \Delta_f(\boldsymbol{x} \mid \{\boldsymbol{x}_{1,j}, ..., \boldsymbol{x}_{i-1,j}\}, \mathcal{S}), \tag{3}$$

where $\mathcal{S}$ represents the observed labeled data from all previous $j-1$ batches, and $\{\boldsymbol{x}_{1,j}, \cdots, \boldsymbol{x}_{i-1,j}\}$ retains the selected $i-1$ samples whose labels are not observed yet within the current j-th batch. This algorithm is theoretically guaranteed to obtain an approximation to the optimal batch-mode active sampling strategy.

2.3 Modeling the Partition Matroid Constraint

Since images of the same patient are very likely to include large pathological information redundancy, we propose to explicitly enforce diversity within the selected images by imposing an additional partition matroid constraint on the original adaptive submodular function maximization problem in Eq. (3).

A partition matroid constraint is defined as follows: $\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_q$ are a partitioning of the set $\mathcal{U}$ if $\mathcal{U} = \bigcup_{1\le i\le q} \mathcal{P}_i$ and $\mathcal{P}_1, \cdots, \mathcal{P}_q$ are disjoint with each other. We require the currently selected batch to include at most one sample from each subset $\mathcal{P}_i$.

More formally, our proposed constrained problem is defined as follows:

$$\begin{aligned} \mathcal{B}^* = \ & \arg\max_{\mathcal{B}\subseteq\mathcal{U}} \Delta_f(\mathcal{B} \mid \mathcal{S}) \\ & \text{subject to } |\mathcal{B}| = k, |\mathcal{B}\cap\mathcal{P}_i| \le 1, k \le q, \forall i \in \{1, ..., q\}, \end{aligned} \tag{4}$$

where $\mathcal{B}^*$ is the optimal k-cardinality batch selected from the current unlabeled dataset $\mathcal{U}$, $\mathcal{P}_1, \cdots, \mathcal{P}_q$ are q disjoint subsets partitioning $\mathcal{U}$, and $\mathcal{S}$ is the set composed of the previously observed labeled data. These disjoint subsets can be obtained through performing clustering according to the structured information of the annotated images.

Within each batch, the i-th sample of the j-th batch is selected as follows

$$\begin{aligned} \boldsymbol{x}^* = \ & \arg\max_{\boldsymbol{x}\in\mathcal{U}} \Delta_f(\boldsymbol{x} \mid \{\boldsymbol{x}_{1,j}, ..., \boldsymbol{x}_{i-1,j}\}, \mathcal{S}) \\ & \text{subject to } cluster(\boldsymbol{x}) \neq cluster(\boldsymbol{x}_{k,j}), \forall k \in \{1, \cdots, i-1\}, \end{aligned} \tag{5}$$

where $cluster(\boldsymbol{x})$ is the index of the cluster that $\boldsymbol{x}$ belongs to.

For the sequential version of this problem, Golovin and Krause[7] have proven that the greedy method can achieve a $(\frac{1}{p+1})$-approximation to the optimum when maximizing f subject to p matroid constraints, which motivates us to generalize this result to the batch mode setting. We propose a practical batch mode active learning algorithm BGAL-PMC, as described in Algorithm 1. In what follows, we show that BGAL-PMC can well approximate the optimal batch selection strategy. Note that $\mathcal{H}$ is the hypothesis set, $\mathcal{H}(\mathcal{S})$ is the reduced hypothesis set which is consistent to the observation $\mathcal{S}$, and $|\mathcal{H}|$ is the size of the hypothesis set.

Algorithm 1. BGAL-PMC (Batch Greedy Active Learning with a Partition Matroid Constraint)

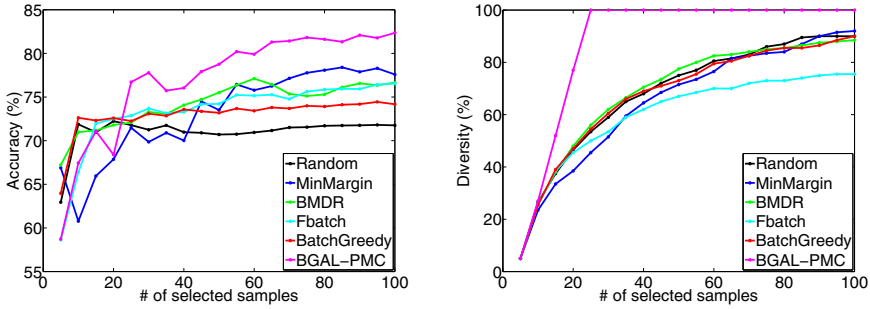
Input: a set of disjoint clusters $\mathcal{P}_1, \mathcal{P}_2, ...\mathcal{P}_q$, previously selected dataset $\mathcal{S}$ and their observed labels $\boldsymbol{y_S}$, unlabeled dataset $\mathcal{U}$, hypothesis set size N, and batch size k.
Ouput: the selected batch $\mathcal{B}$ and their labels $\boldsymbol{y_B}$.
Sample a hypothesis set $\mathcal{H} = \{h_1, h_2, ...h_N\}$ using $\boldsymbol{y_S}$;
initialize $\mathcal{B} = \emptyset$, $D = \emptyset$, and $\mathcal{T} = \emptyset$;
for $i = 1$ **to** k **do**
  for $j = 1$ **to** $|\mathcal{U}|$ **do**
    $score(\boldsymbol{x}_j) = \mathbb{E}_{y \in \{-1,1\}}[|\mathcal{H}(\{(\boldsymbol{x}, y) \mid \boldsymbol{x} \in \mathcal{B} \cup \{\boldsymbol{x}_j\}\})|]$
  end for
  while true **do**
    $\boldsymbol{x}^* = arg\min_{\boldsymbol{x} \in \mathcal{U} \setminus \{\mathcal{B} \cup \mathcal{T}\}} score(\boldsymbol{x})$
    $ind = cluster(\boldsymbol{x}^*)$
    if $ind \notin D$ **then**
      $\mathcal{B} = \mathcal{B} \cup \{\boldsymbol{x}^*\}$, $D = D \cup \{ind\}$
      break
    else
      $\mathcal{T} = \mathcal{T} \cup \{\boldsymbol{x}^*\}$
    end if
  end while
end for
query the labels $\boldsymbol{y_B}$ for $\mathcal{B}$.

Theorem 1. *Given a monotonic and submodular function f and a label distribution P such that (f, P) is adaptive submodular, when maximizing f subject to a partition matroid constraint, the expected cost of the BGAL-PMC algorithm is at most* $2(\ln(|\mathcal{H}| - 1) - 1)$ *times the expected cost of the optimal batch selection strategy.*

The proof of Theorem 1 is provided in the supplemental material. This theorem guarantees that BGAL-PMC needs at most $2(\ln(|\mathcal{H}| - 1) - 1)$ times more batches than those required by the optimal batch selection strategy. Note that directly searching for the optimal selection strategy takes exponential time. To sample a finite hypothesis set $\mathcal{H}$, we employ the hit-and-run sampler [11] to generate a set of linear separators, which has been used by [2] and proven effective for active learning problems.

3 Experiments

Experimental Settings: Our experiments are conducted on a large database of histopathological images from breast microscopic tissues [4,17]. This database contains more than two thousand images, gathered from around a hundred patients. Each image is labeled as benign category (usual ductal hyperplasia (UDH)) or actionable category (atypical ductal hyperplasia (ADH) and ductal carcinoma in situ (DCIS)) by pathologists, which are development procedures from a normal terminal duct-lobular unit to an invasive cancer. Classifying these two categories is an important clinical problem since the therapy planning and management relies on the diagnosis of UDH



(a) Accuracy Comparison (b) Diversity Comparison

Fig. 1. (a) Learning curves of the proposed BGAL-PMC and other 5 methods on the breast microscopic tissues image dataset. X-axis is the number of selected images while Y-axis is the accuracy as the number of selected training images increases. BGAL-PMC (the pink curve) outperforms the other 5 methods significantly; (b) The diversity curves of all 6 methods. X-axis is the number of selected images while Y-axis is the diversity of the selected set as the number of selected images increases. Note that the diversity here is defined as the perentage of partitioning clusters being covered.

and ADH/DCIS. It is also very challenging due to the subtle differences between categories. High-dimensional (i.e., 10000) texture features are extracted from each image. We randomly split the dataset into 50% training to actively select candidate images and 50% testing to test the learned classifier. We also ensure that images for a particular patient are either in the training set or in the testing set. We randomly split 10 times and the average performance is reported.

Five active learning methods are compared, i.e., Random Selection, Min Margin [15], Fbatch [10], BMDR [16], and BatchGreedy [2]. Note that the Random Selection is equivalent to the passive learning setting. In our method, we partitioned the dataset into 20 disjoint subsets using both the structured information and image texture features by K-means. Since it's difficult and time-consuming to sample hyperplanes uniformly in high dimensional space, we follow [2] to reduce the dimension to 100 to sample the hypothesis set $\mathcal{H}$. For fair comparison, we use SVM classifier for all methods, with the same parameters tuned via five-fold cross validation. We set batch size at 5 throughout the experiments. Two positive images and two negative images are randomly selected for initialization. The size of the hypothesis set is set at 300, which is empirically large enough in our experiments. All experiments are conducted on a 2.80GHz i7 CPU with 16 cores and 16G RAM in Matlab implementation.

Results: Fig. 1(a) shows the classifier learning curves as selected samples increase. Not surprisingly, all five active learning methods perform better than random selection, which manifest the effectiveness of active learning. In particular, the proposed BGAL-PMC performs significantly better than all other four active learning methods. Min Margin method as a classical active learning baseline is the second-best in our experiments. Although Fbatch, BMDR and BatchGreedy perform well in the first 20 selected samples, the improvement of their accuracy is less substantial when more

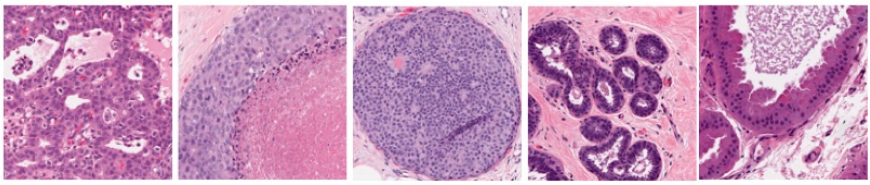


Fig. 2. One example batch of selected images using our proposed method. The first 3 are actionable, and the last 2 are benign. 5 images are selected from distinct clusters.

Table 1. Comparsion of the average time to select a single batch of images for 5 active learning algorithms (batch size=5)

Methods	MinMargin[15]	BMDR[16]	FBatch[10]	BatchGreedy[2]	BGAL-PMC
Time (seconds)	3.13	17.63	128.13	1.97	1.98

batches are selected. The reason is that all other methods do not take the information of clusters into consideration. Therefore, their selected images may include information redundancy, which downgrades their performances. On the other hand, trivially using cluster information cannot achieve the same accuracy either. We tested samping from randomly-chosen distinct clusters, as an alternative baseline. It achieved 77% accuracy when selecting 100 samples which is better than some baselines, but is still significantly worse than our proposed method. Leveraging image structured information may be a general paradigm to boost active learning performance, but our proposed matroid constraint is a more effective and theoretical sound method. With less than 5% data labeled, our method achieves 83% prediction accuracy. This accuracy is at least 6% higher than all compared methods. In fact, when 80% data is labeled, the prediction accuracy is 87%, which is merely 4% higher than our method but use much more labeled samples than us. Therefore, this scheme considerably reduces the label effort from pathologists, without significantly sacrificing the accuracy.

We further investigated the diversity of all methods, as shown in Fig. 1(b). The diversity here is defined as the coverage rate of the clusters. Since we enforce the partition matroid constraint explicitly, BGAL-PMC covered all the clusters in much fewer iterations than other methods. Fig. 2 is one selected batch using our proposed method, to show the diversity of our selections visually. We also compared the running time, as shown in Table 1. In our experiments, BatchGreedy and BGAL-PMC are much more scalable than other active learning algorithms. BatchGreedy is slightly faster than ours (1.97s vs. 1.98s), both of which are negligible in the practical use of active learning.

4 Conclusion

In this paper, we proposed a novel batch mode active learning approach which leverages the structured information of annotated histopathological images. We formulated the batch mode active learning problem as a submodular function maximization problem with a partition matroid constraint, which prompts us to design an efficient greedy

algorithm for approximate combinatorial optimization. We further provided a theoretic bound characterizing the quality of the solution achieved by our algorithm. We compared the proposed active learning approach against several state-of-the-art active learning methods on a large database of histopathological images, and demonstrated the superiority of our approach in performance. The spirit of our active learning method capitalizing on submodular optimization is generic, and can thus be applicable to other problems in medical image analysis. In the future, we will also explore more sophiscated ways to extract structured infomation.

References

1. Balcan, M.F., Hanneke, S., Vaughan, J.W.: The true sample complexity of active learning. Machine Learning 80(2-3), 111–139 (2010)
2. Chen, Y., Krause, A.: Near-optimal batch mode active learning and adaptive submodular optimization. In: Proc. ICML (2013)
3. Doyle, S., Agner, S., Madabhushi, A., Feldman, M., Tomaszewski, J.: Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In: Proc. ISBI (2008)
4. Dundar, M.M., Badve, S., Bilgin, G., Raykar, V., Jain, R., Sertel, O., Gurcan, M.N.: Computerized classification of intraductal breast lesions using histopathological images. IEEE Transactions on Biomedical Engineering 58(7), 1977–1984 (2011)
5. Fisher, M.L., Nemhauser, G.L., Wolsey, L.A.: An analysis of approximations for maximizing submodular set functions–ii. In: Polyhedral Combinatorics pp. 73–87 (1978)
6. Foran, D.J., Yang, L., et al.: Imageminer: a software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology. JAMIA 18(4), 403–415 (2011)
7. Golovin, D., Krause, A.: Adaptive submodular optimization under matroid constraints. arXiv preprint arXiv:1101.4450 (2011)
8. Golovin, D., Krause, A.: Adaptive submodularity: Theory and applications in active learning and stochastic optimization. JAIR 42(1), 427–486 (2011)
9. Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B.: Histopathological image analysis: A review. IEEE Reviews in Biomedical Engineering 2, 147–171 (2009)
10. Hoi, S.C., Jin, R., Zhu, J., Lyu, M.R.: Batch mode active learning and its application to medical image classification. In: Proc. ICML (2006)
11. Lovász, L.: Hit-and-run mixes fast. Mathematical Programming 86(3), 443–461 (1999)
12. Petushi, S., Garcia, F.U., Haber, M.M., Katsinis, C., Tozeren, A.: Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. BMC Medical Imaging 6(1), 14 (2006)
13. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2002)
14. Settles, B.: Active learning literature survey. Technical Report, University of Wisconsin, Madison (2010)
15. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. Journal of Machine Learning Research 2, 45–66 (2002)
16. Wang, Z., Ye, J.: Querying discriminative and representative samples for batch mode active learning. In: Proc. KDD (2013)
17. Zhang, X., Liu, W., Zhang, S.: Mining histopathological images via hashing-based scalable image retrieval. In: ISBI (2014)

Unsupervised Unstained Cell Detection by SIFT Keypoint Clustering and Self-labeling Algorithm

Firas Mualla[1], Simon Schöll[1,3,4], Björn Sommerfeldt[2], Andreas Maier[1,3], Stefan Steidl[1], Rainer Buchholz[2], and Joachim Hornegger[1,3]

[1] Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg
[2] Institute of Bioprocess Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg
[3] SAOT Graduate School in Advanced Optical Technologies
[4] ASTRUM IT GmbH

Abstract. We propose a novel unstained cell detection algorithm based on unsupervised learning. The algorithm utilizes the scale invariant feature transform (SIFT), a self-labeling algorithm, and two clustering steps in order to achieve high performance in terms of time and detection accuracy. Unstained cell imaging is dominated by phase contrast and bright field microscopy. Therefore, the algorithm was assessed on images acquired using these two modalities. Five cell lines having in total 37 images and 7250 cells were considered for the evaluation: CHO, L929, Sf21, HeLa, and Bovine cells. The obtained F-measures were between 85.1 and 89.5. Compared to the state-of-the-art, the algorithm achieves very close F-measure to the supervised approaches in much less time.

1 Introduction

Cell detection plays a vital role in biomedical image analysis. Automatic image-based cell detection approaches can be used for estimating the number of cells [1,2], initializing cell segmentation algorithms [3], cell tracking [4], and for extracting features which can be used for further analysis such as cell viability determination [5].

In fluorescence microscopy, cells are stained using a fluorescent dye. This reshapes the cell detection problem as a relatively easy task due to the high contrast obtained by staining. On the other hand, in some biological applications [6], it is desired to avoid staining because of its side effects on cells. In this case, cell detection is more challenging and sometimes very difficult [7,8,9].

We know from bioprocess engineers that, for unstained cell imaging, bright field and phase contrast are the most widely used microscopic modalities. Therefore, they form together a very appropriate choice for the evaluation of unstained cell detection.

Several machine learning approaches have been proposed in the literature in order to cope with the difficulty of the problem. Some approaches follow a pixel-wise classification strategy [5,8,10,11]. Others perform the classification at the level of image interest points [9,12,13,14]. The latter have some useful properties. First, the problem is sparse compared to the pixel-wise classification. Second, the interest points are characterized by features and/or descriptors which can be utilized for detection as in [9, 12]. Third, they can be employed to achieve scale- and orientation-invariant training as in [9].

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 377–384, 2014.

All the previous approaches are dependent on supervised learning. The latter transfers part of its inductive bias to the training data which makes the approach adaptable by simply changing the training set. This has the advantage that it can model very complicated situations and provide reliable results as long as the training set is representative. On the other hand, its drawback is that it requires labeled ground truth. In many cases, the users of cell image analysis software would sacrifice some detection accuracy in favor of having a labeling-free system. This preference becomes more serious when the system has to be trained for each new cell line.

In this paper, we advocate an unsupervised machine learning approach for unstained cell detection. Technically speaking, we also employ supervised learning, but with ground truth learned automatically from the input image. The proposed approach was tested on five cell lines with diverse visual appearance. Our results show that we are very close in terms of detection rate to the state-of-the-art supervised learning approaches. However, our approach has a much faster runtime and does not require manually-labeled ground truth.

Related Work. In [12] and [14] on phase contrast microscopy and [9] on bright field microscopy, the training was done at two levels: First, cells and background are separated by machine learning techniques. Second, the difference between interest points belonging to the same cell and interest points belonging to neighboring cells is learned. The support vector machine (SVM) classifier and interest points detected by a set of Laplacian filters were utilized in [12] and [14] while the random forest classifier and the scale invariant feature transform (SIFT) keypoints were used in [9]. The previous approaches require ground truth of segmented cells. In other words, cell borders should be delineated and each cell should have a distinguishing identifier in the ground-truth mask. In [13], maximally stable extremal regions (MSER) keypoints were utilized and a structured SVM was used to learn a bijective mapping between the MSER regions and the ground-truth cell centers. Compared to [9], [12], and [14], this approach has the advantage that it is easier to train because only cell centers are required as ground truth. The closest to our approach is [9], but ours uses unsupervised learning and it is thus labeling-free.

2 Methods

We make a heavy use of SIFT related concepts. Therefore, we introduce SIFT in Section 2.1 and we then describe our method in Section 2.2.

2.1 SIFT

SIFT [15] is a local image feature detector and descriptor. Each detected keypoint is characterized by its spatial coordinates, a scale, an orientation, a difference of Gaussians (DOG) value, and a principal curvatures ratio (PCR) value. The DOG value indicates the keypoint strength and its range is proportional to the dynamic image range. Its sign is positive for black-on-white blobs and negative for white-on-black blobs. The PCR value is defined as [15]:

$$\mathrm{PCR} = \frac{\mathrm{Tr}(\mathbf{H}_{DOG})^2}{\mathrm{Det}(\mathbf{H}_{DOG})} \tag{1}$$

where $\mathbf{H}_{DOG}$ is the Hessian of DOG, Tr is the trace, and Det is the determinant. PCR has a minimum of 4 for the isotropic blobs and its value increases theoretically until $+\infty$ by increased blob anisotropy.

2.2 Cell Detection by Keypoint Clustering and Self-labeling

2.2.1 Keypoint Extraction

Our algorithm starts by extracting SIFT keypoints of the input image $\mathbf{I}$. These keypoints are not thresholded using the PCR or the DOG values. In other words, all detected SIFT keypoints of all strength and anisotropy values are considered at this step.

2.2.2 Blob Type Detection

The keypoint blob type is determined by the DOG sign. As mentioned in Section 2.1, it is either black on white (+1), or white on black (-1). Based on [9], the blob type is computed using the following equation:

$$\beta = \text{sign}\left(\frac{\sum_{i=1}^{N} \omega_i \left|DOG(\mathbf{p}_i)\right| H(DOG(\mathbf{p}_i))}{\sum_{i=1}^{N} \omega_i \left|DOG(\mathbf{p}_i)\right|} - \frac{1}{2}\right) \tag{2}$$

$$\omega_i = \frac{s(\mathbf{p}_i)}{\text{PCR}(\mathbf{p}_i)} \tag{3}$$

where $\mathbf{p}_i$, $i = 1..N$ are the keypoints in the image as obtained in Step 2.2.1, N is their number, $s(\mathbf{p}_i)$ is the scale of $\mathbf{p}_i$, and H is the Heaviside step function. If $\beta = +1$, the positive DOG keypoints are considered while the negative DOG keypoints are discarded, and vice versa.

2.2.3 Scale Adaptive Smoothing

The image $\mathbf{I}$ is smoothed with a Gaussian kernel whose standard deviation is the mean keypoint scale. The latter is computed using the following equation:

$$\sigma = \frac{\sum_{i=1}^{M} \left|DOG(\mathbf{p}_i)\right| s(\mathbf{p}_i)}{\sum_{i=1}^{M} \left|DOG(\mathbf{p}_i)\right|} \tag{4}$$

where M is the number of the keypoints resulting from Step 2.2.2, i.e. only one blob type is considered.The smoothed image $\mathbf{I}_\sigma$ is saved for further processing.

2.2.4 Second Keypoint Extraction

Step 2.2.1 is applied on the smoothed image $\mathbf{I}_\sigma$ and the keypoints which conform to the previously computed β are considered while the others are discarded.

2.2.5 Cell/Background Keypoint Clustering

At this step, the keypoints are clustered into one of two categories: cells and background. K-medians clustering, i.e. intra-cluster $\ell 1$-norm minimization, is applied with $K = 2$. One-dimensional Otsu thresholding is applied on the DOG values of the keypoints and the two resulting clusters are used to initialize the Lloyd's iteration. The features are modality-specific. For bright field microscopy, at each keypoint $\mathbf{p}_i$, we employ $\text{DOG}(\mathbf{p}_i)$ and smoothed image intensity $\mathbf{I}_\sigma(\mathbf{p}_i)$ as features. For phase contrast

microscopy, we use $\mathrm{DOG}(\mathbf{p}_i)$ and $\mathrm{VAR}(\mathbf{I}, \mathbf{p}_i, \sigma)$. The latter is the local variance of the original image $\mathbf{I}$ within a square neighborhood centered at $\mathbf{p}_i$ with a half side-length equal to σ (up to an integer approximation). The features are normalized to $[0, 1]$ so that they contribute equally to the $\ell 1$-norm. After termination, the keypoints which belong to the background cluster are discarded.

2.2.6 Cell/cell Keypoint Clustering

The goal of this step is to cluster the cell keypoints resulting from the previous step into N_c clusters where two keypoints belong to the same cluster if and only if they belong to the same cell. N_c is not known a priori. In order to achieve this goal, a classifier which ranks each pair of keypoints as belonging to the same cell or not is required [9], [14]. We propose to learn this classifier from the input image using a self-labeling algorithm instead of manually-labeled ground truth. Informally speaking, the algorithm trains a keypoint-pair classifier on *extreme* cases (for which ground truth labels can be assumed) and applies the resulting classifier on *intermediate* cases. This is achieved as follows:

1) Consider Ψ to be a set of keypoint pairs defined as: $(\mathbf{p}_i, \mathbf{p}_j) \in \Psi \Leftrightarrow \|\mathbf{p}_i - \mathbf{p}_j\|_2 \geq \rho$ where $\rho = \alpha \cdot \sigma$ and α is a constant. ρ must be larger than the maximum cell length. Due to the use of SIFT, *safe* values for α can be set easily regardless of the image resolution or cell type. We set it to 10 in our experiments.
2) Randomly choose N_1 elements, i.e. keypoint pairs, from Ψ. Label each of them as *cross* which means that the two corresponding keypoints belong to two different cells.
3) Randomly choose N_2 keypoints from the set of cell keypoints and form the set Ω. The probability of selecting a keypoint is proportional to its scale. Both N_1 and N_2 were set to 100 in our experiments.
4) Motivated by the intuition that *short* line segments are very unlikely to span two cells: For each element $\mathbf{p}_i$ in Ω, choose a random orientation θ_i and form the point $\mathbf{q}_i = \mathbf{p}_i + (s(\mathbf{p}_i)\cos(\theta_i), s(\mathbf{p}_i)\sin(\theta_i))$. Label each pair $(\mathbf{p}_i, \mathbf{q}_i)$ as *inner* which means that the two corresponding points belong to the same cell. The labels obtained by this step and by Step 2 are illustrated in Figure 1 (b).
5) For each *inner*/*cross* pair $(\mathbf{p}_i^*, \mathbf{p}_j^*)$, extract the following feature after [9]: $F_{ij} = \mathbf{I}_\sigma(\mathbf{p}_i^*) - 2\,\mathrm{extremum}_{ij} + \mathbf{I}_\sigma(\mathbf{p}_j^*)$. $\mathrm{extremum}_{ij}$ is, by definition, either the maximum (when $\beta = +1$) or the minimum ($\beta = -1$) intensity along the line segment between $\mathbf{p}_i^*$ and $\mathbf{p}_j^*$.
6) Estimate the two class conditional densities $\mathrm{P}(F|\textit{inner})$ and $\mathrm{P}(F|\textit{cross})$ assuming a Gaussian distribution.

So far, a keypoint-pair classifier was trained using the input image. The *posterior* probability $\mathrm{P}(\textit{cross}|F)$, assuming equal priors, is then used to rank each two nearby keypoints (cf. Figure 1 (c)). This ranking expresses the probability that they belong to two different cells. In order to reduce runtime, only the three nearest neighbors of each keypoint are considered. The resulting ranks are then used as input for an agglomerative hierarchical clustering with average linkage similar to [9]. The resulting clusters at a cut-off equal to 0.5 (cf. Figure 1 (d)) represent the detected cells. Inside each cluster, the arithmetic average of the keypoint coordinates identifies the center of a detected cell.

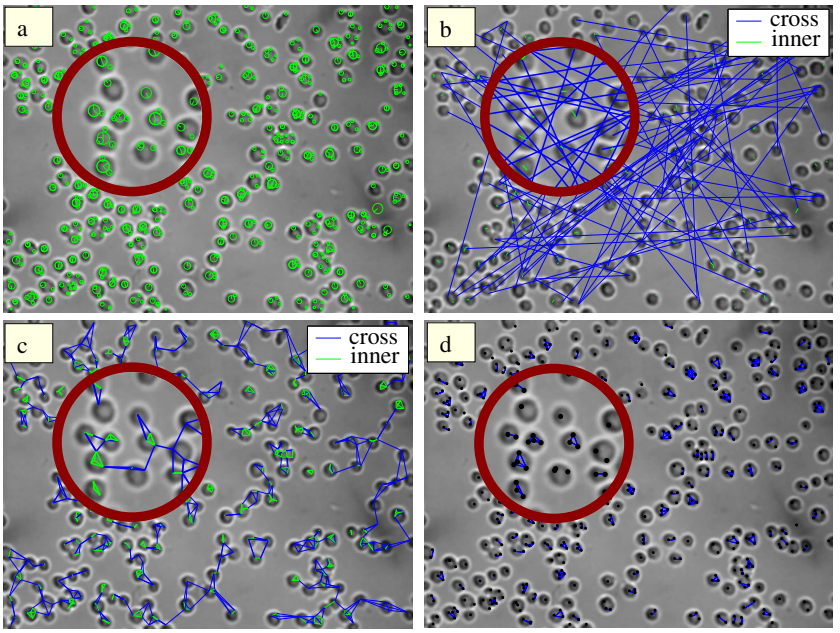


Fig. 1. Illustration of the cell/cell keypoint clustering. The circle inside each figure shows a magnified view. a) Cell keypoints resulting from the cell/background K-medians clustering. b) Point pairs chosen by the self-labeling algorithm for training a cell boundary potential. Each pair is indicated by a line segment. c) The learned boundary potential is employed to rank nearby keypoint pairs. The output is probabilistic, but only the binary classification result is shown. d) Result of hierarchical clustering using the ranks obtained from the previous step. Each cluster represents a detected cell.

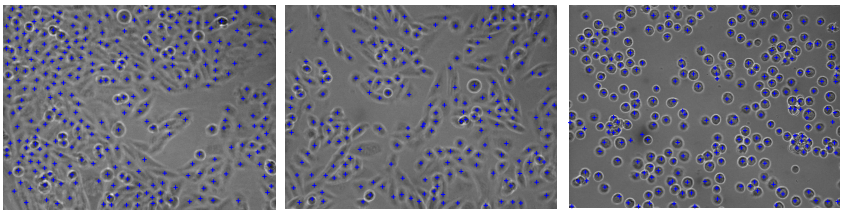
3 Evaluation

Table 1 contains a summary of the datasets used in the evaluation. The ground-truth type of all datasets except [13] is cell border delineation, while in the dataset of [13] a dot is marked at the center of each cell. This difference in ground-truth representation leads to a difference in the evaluation procedure. In all datasets except [13], a cell is considered detected if the hit point belongs to the cell mask and the *centeredness error* is used to assess the deviation from the cell center. Centeredness error is defined after [9] as the distance between the hit point and the cell's center of mass normalized by the cell major axis length and averaged over all correctly-detected cells in the considered image. In the dataset of [13], cell masks are not available. Therefore, a cell is considered detected if the distance to the ground-truth cell center is less than the minimum cell radius. The latter was set after [13] to 5 pixels. Figure 2 exemplifies detection results of our approach for all datasets. Quantitative evaluation and comparison with the state-of-the-art are described in the next paragraph. The evaluation results of all approaches except ours are given according to their corresponding papers: [9], [14], and [13].

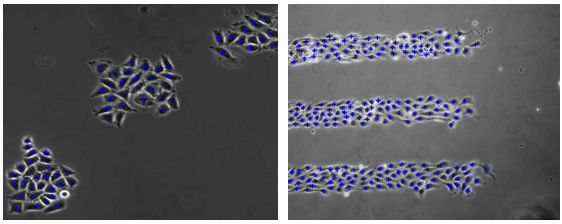
A comparison with [9] on bright field microscopy is shown in Table 2. The figures of [9] in Table 2 were obtained by image-wise cross-validation in each cell line: One

image per cell line is used for training and the other images of the same cell line are used for testing. The results of our approach were obtained by averaging each of the F-measure, time, and centeredness error over images per cell line. A comparison with [13] and [14] on phase contrast microscopy is shown in Table 3. The shown results of the approaches [13] and [14] in Table 3 were generated by the hold-out method: [13] was trained using 11 images and tested on other 11 images. Similarly, [14] was trained using 10 images and tested on other 10 images. We evaluated our approach on the same images which were used for *testing* each of them (the images described in Table 1). Tables 2 and 3 show that the proposed approach is very close in terms of F-measure and centeredness error (when available) to the supervised approaches. However, our approach is much faster especially when compared with the phase contrast approaches where it is one or two orders of magnitude faster.

The blob type was correctly picked for all images by Eq. 2. As can be seen in Eq. 2, this blob type is decided by the sign function. Therefore, the reliability of the decision is proportional to the absolute value of the sign operand. We observed a little improvement (data not shown) of this reliability when both PCR and scale are used for weighting (as in Eq. 3) compared to the case when only the scale is used.



(a) CHO (b) L929 (c) Sf21



(d) HeLa (e) Bovine

Fig. 2. Samples of the detection results. Each plus sign marks a detected cell.

Table 1. Datasets used in the evaluation

Cell line	Modality	Resolution	#Images	#Cells	Ground truth	Source
CHO	bright field	1280×960	6	1431	border delineation	[9]
L929	bright field	1280×960	5	1078	border delineation	[9]
Sf21	bright field	1280×960	5	1001	border delineation	[9]
HeLa	phase contrast	400×400	11	1156	center dots	[13]
Bovine	phase contrast	680×512	10	2584	border delineation	[14]

Table 2. Comparison with the state-of-the-art on bright field microscopy

	F-measure (%)			Time (seconds)			Centeredness error		
	CHO	L929	Sf21	CHO	L929	Sf21	CHO	L929	Sf21
Supervised Mualla et al. [9]	84.2	86.5	97.0	45.9	36.7	40.7	0.48	0.38	0.16
Proposed approach	85.1	88.3	89.5	10.5	10.9	14.4	0.40	0.42	0.23

Table 3. Comparison with the state-of-the-art on phase contrast microscopy

	F-measure (%)		Time (seconds)		Centeredness error	
	HeLa	Bovine	HeLa	Bovine	HeLa	Bovine
Supervised Pan et al. [14]	-	94.6	-	900.0	-	-
Supervised Arteta et al. [13]	88.0	-	30.0	-	-	-
Proposed approach	88.7	86.0	1.5	3.5	-	0.11

4 Discussion and Conclusion

Both blob type detection and scale adaptive smoothing were proposed in the supervised approach of [9]. In contrast to [9], where only keypoints which belong to cells (known from ground truth) are considered, we compute the blob type β in an unsupervised manner by considering all keypoints. In addition, we use both scale and PCR to weigh the keypoint contribution to β whereas only scale is used in [9]. For the scale adaptive smoothing, we use a weighted average instead of the simple arithmetic average used in [9]. In general, we can conclude that SIFT can be successfully employed for *unsupervised* structure-of-interest measurements such as mean scale and dominant curvature direction.

In the cell/cell clustering step, a self-labeling algorithm was employed to train a ranking classifier. This classifier learns from extreme cases and applies the learned model on intermediate ones. In other words, training and testing feature vectors are drawn from different distributions. Therefore, the features should be chosen carefully so that they do not overfit the training samples. With this in mind, we confined ourselves to use a one-dimensional feature space and a simple generative model. In future, we plan to improve the cell/cell clustering by applying transductive transfer learning techniques. On the other hand, for the possibly less-reliable cell/background clustering, we think that applying transductive learning methods may alleviate the limitations of K-medians. In the self-labeling algorithm, due to the use of SIFT, it was possible to define a scale-invariant notion of the *extreme* cases. Consequently, the algorithm could successfully detect cells in images of different resolutions and/or cell types without any change in the parameter values.

The proposed approach achieves detection accuracy which is very close to three state-of-the-art supervised cell detection approaches in much less time, without training data, and without manual parameter-tuning. We thus believe that the cell detection problem is, to a large extent, solvable by self-supervised techniques which learn from the input image itself.

Acknowledgment. The authors gratefully acknowledge funding of the Erlangen Graduate School in Advanced Optical Technologies (SAOT) by the German Research

Foundation (DFG) in the framework of the German excellence initiative. Special thanks go to Mr. Arteta and Dr. Pan for providing us with the two phase contrast datasets.

References

1. Sjöström, P.J., Frydel, B.R., Wahlberg, L.U.: Artificial neural network-aided image analysis system for cell counting. Cytometry 36(1), 18–26 (1999)
2. Loukas, C.G., Wilson, G.D., Vojnovic, B., Linney, A.: An image analysis-based approach for automated counting of cancer cell nuclei in tissue sections. Cytometry Part A 55A(1), 30–42 (2003)
3. Ali, R., Gooding, M., Szilágyi, T., Vojnovic, B., Christlieb, M., Brady, M.: Automatic segmentation of adherent biological cell boundaries and nuclei from brightfield microscopy images. Machine Vision and Applications 23(4), 607–621 (2012)
4. Li, K., Chen, M., Kanade, T., Miller, E., Weiss, L., Campbell, P.: Cell population tracking and lineage construction with spatiotemporal context. Medical Image Analysis 12(5), 546–566 (2008)
5. Long, X., Cleveland, W., Yao, Y.: Automatic detection of unstained viable cells in bright field images using a support vector machine with an improved training procedure. Computers in Biology and Medicine 36(4), 339–362 (2006)
6. Lulevich, V., Shih, Y.P., Lo, S.H., Liu, G.Y.: Cell tracing dyes significantly change single cell mechanics. The Journal of Physical Chemistry B 113(18), 6511–6519 (2009)
7. van Opstal, W., Ranger, C., Lejeune, O., Forgez, P., Boudin, H., Bisconte, J., Rostene, W.: Automated image analyzing system for the quantitative study of living cells in culture. Microscopy Research and Technique 28(5), 440–447 (1994)
8. Long, X., Cleveland, W., Yao, Y.: A new preprocessing approach for cell recognition. IEEE Transactions on Information Technology in Biomedicine 9(3), 407–412 (2005)
9. Mualla, F., Schöll, S., Sommerfeldt, B., Maier, A., Hornegger, J.: Automatic cell detection in bright-field microscope images using SIFT, random forests, and hierarchical clustering. IEEE Transactions on Medical Imaging 32(12), 2274–2286 (2013)
10. Nattkemper, T., Ritter, H., Schubert, W.: Extracting patterns of lymphocyte fluorescence from digital microscope images. In: Intelligent Data Analysis in Medicine and Pharmacology, pp. 79–88 (1999)
11. Mualla, F., Schöll, S., Sommerfeldt, B., Maier, A., Steidl, S., Buchholz, R., Hornegger, J.: Using the low-pass monogenic signal framework for cell/background classification on multiple cell lines in bright-field microscope images. International Journal of Computer Assisted Radiology and Surgery, 1–8 (2013)
12. Pan, J., Kanade, T., Chen, M.: Learning to detect different types of cells under phase contrast microscopy. In: Microscopic Image Analysis with Applications in Biology (2009)
13. Arteta, C., Lempitsky, V., Noble, J.A., Zisserman, A.: Learning to detect cells using non-overlapping extremal regions. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part I. LNCS, vol. 7510, pp. 348–356. Springer, Heidelberg (2012)
14. Pan, J., Kanade, T., Chen, M.: Heterogeneous conditional random field: Realizing joint detection and segmentation of cell regions in microscopic images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2940–2947 (2010)
15. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)

Selecting Features with Group-Sparse Nonnegative Supervised Canonical Correlation Analysis: Multimodal Prostate Cancer Prognosis

Haibo Wang*, Asha Singanamalli, Shoshana Ginsburg, and Anant Madabhushi

Department of Biomedical Engineering, Case Western Reserve University, OH 44106 USA
{hxw285,axs1018,sbg52,anantm}@case.edu

Abstract. This paper presents Group-sparse Nonnegative supervised Canonical Correlation Analysis (GNCCA), a novel methodology for identifying discriminative features from multiple feature views. Existing correlation-based methods do not guarantee positive correlations of the selected features and often need a pre-feature selection step to reduce redundant features on each feature view. The new GNCCA approach attempts to overcome these issues by incorporating (1) a nonnegativity constraint that guarantees positive correlations in the reduced representation and (2) a group-sparsity constraint that allows for simultaneous between- and within- view feature selection. In particular, GNCCA is designed to emphasize correlations between feature views and class labels such that the selected features guarantee better class separability. In this work, GNCCA was evaluated on three prostate cancer (CaP) prognosis tasks: (i) identifying 40 CaP patients with and without 5-year biochemical recurrence following radical prostatectomy by fusing quantitative features extracted from digitized pathology and proteomics, (ii) predicting *in vivo* prostate cancer grade for 16 CaP patients by fusing T2w and DCE MRI, and (iii) localizing CaP/benign regions on MR spectroscopy and MRI for 36 patients. For the three tasks, GNCCA identifies a feature subset comprising 2%, 1% and 22%, respectively, of the original extracted features. These selected features achieve improved or comparable results compared to using all features with the same Support Vector Machine (SVM) classifier. In addition, GNCCA consistently outperforms 5 state-of-the-art feature selection methods across all three datasets.

1 Introduction

Availability of multiple data streams presents an opportunity to fuse and combine multimodal biomarkers, for potentially improving performance of predictors of disease diagnosis and prognosis. Canonical correlation analysis (CCA) [1] addresses this multimodal fusion task by attempting to maximize the correlations of the multiple data sources. Supervised multi-view CCA (SMVCCA) [2] combines the principle of CCA

* Research was supported by R01CA136535-01, R01CA140772-01, R21CA167811-01, R01DK098503-02, PC120857, and the Ohio Third Frontier Technology development Grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 385–392, 2014.

and linear discriminant analysis (LDA), to find a subspace that maximizes the correlations of multi-view signals and ensures discriminability of provided class labels. Despite its advantages, SMVCCA has some key limitations. First, the CCA components can be negatively correlated, which is less interpretable in practice and hurts the positive dependency between data and their class labels. Secondly, a pre-feature selection step is required in order to reduce redundant features [2]. Thirdly, SMVCCA emphasizes the correlations of all modalities, but neglects modality-specific information.

In this paper, we present Group-sparse Nonnegative supervised CCA (GNCCA), which incorporates (1) nonnegativity and (2) group sparsity constraints to overcome the aforementioned issues of SMVCCA. The nonnegativity is applied onto both the projection and coefficient factor matrices, thus ensuring latent components are positively correlated. Group sparsity allows to simultaneously perform view fusion and within-view feature selection, thus capturing both view-shared and view-specific information. In particular, in order to encourage the association between feature views and class labels, GNCCA keeps the projection section of the label view non-sparse. Although sparse nonnegative CCA has been previously proposed [3], the framework can only calculate the projection of a single view at a time, which has no group sparsity and is difficult to be extended for multiple views. In contrast, GNCCA is based on a more general matrix factorization scheme, which can update all view projections simultaneously.

By ranking the original features based on their values in the projected subspaces [4], GNCCA can be easily used as a multi-modal feature selector, providing an efficient way to interpret the importance of each of the original features for feature-based classification tasks. Unlike the existing feature selection tools, which either neglect view information [4–7] or address only class separability with group lasso [8], GNCCA considers both view association and the discriminability of the selected features.

In this work, we compared GNCCA with 5 related feature selection schemes on three prostate cancer (CaP) prognosis tasks: (i) identifying 40 CaP patients with and without 5-year biochemical recurrence following radical prostatectomy by fusing quantitative features extracted from digitized pathology and proteomics, (ii) predicting *in vivo* prostate cancer grade for 16 CaP patients by fusing T2w and DCE MRI, and (iii) localizing CaP/benign regions on MR spectroscopy and MRI for 36 patients.

2 Related Work

In this section we first describe supervised multi-view CCA (SMVCCA) and then discuss previous work in sparse nonnegative CCA.

Supervised Multi-View Canonical Correlation Analysis (SMVCCA). Provided n data samples $\mathbf{X}$ from K feature views, multi-view CCA [1] seeks a set of linear transformations $\{\mathbf{W}_1, \mathbf{W}_2, ..., \mathbf{W}_K\}$ such that the sum of the correlations of every two views are maximized. In order to ensure class separability, supervised multi-view CCA (SMVCCA) [2] was proposed to simultaneously maximize the correlations of feature views and the correlations between $\mathbf{X}$ and its class labels $\mathbf{Y}$, which yields:

$$\arg \max_{\mathbf{W}_x, \mathbf{W}_y} trace(\mathbf{W}_x^T \mathbf{C} \mathbf{W}_x) + 2trace(\mathbf{W}_x^T \mathbf{X} \mathbf{Y}^T \mathbf{W}_{(y)}) \tag{1}$$

$$s.t.\ \ \mathbf{W}^T \mathbf{C}_{d_{xy}} \mathbf{W} = \mathbf{I},\ \mathbf{W}_{(1)}^T \mathbf{C}_{d_{xy}}^{(11)} \mathbf{W}_{(1)} = \ldots = \mathbf{W}_{(K)}^T \mathbf{C}_{d_{xy}}^{(KK)} \mathbf{W}_{(K)} = \mathbf{W}_{(y)}^T \mathbf{Y} \mathbf{Y}^T \mathbf{W}_{(y)},$$

where $\mathbf{W} = [\mathbf{W}_{(1)}^T, \cdots, \mathbf{W}_{(K)}^T, \mathbf{W}_{(y)}^T]^T$, $\mathbf{I}$ is an identity matrix, $\mathbf{C}$ is the covariance matrix of the K views of $\mathbf{X}$, $\mathbf{C}_{d_{xy}}$ is the block-diagonal section of $\mathbf{C}$, and $\mathbf{C}_{d_{xy}}^{(jj)}$ is the j^{th} diagonal block of $\mathbf{C}_{d_{xy}}$ [2].

Sparse Nonnegative CCA. Sparse nonnegative CCA, which was first proposed in [3], takes the form:

$$\begin{aligned} &\max_{\mathbf{w}_1,\mathbf{w}_2} \ \mathbf{w}_1^T \mathbf{C}_{12} \mathbf{w}_2, \\ &s.t. \ \ \|\mathbf{w}_1\|^2 \le 1, \ \|\mathbf{w}_2\|^2 \le 1, \ P(\mathbf{w}_1) \le c_1, \ P(\mathbf{w}_2) \le c_2, \ \mathbf{w}_1 \succeq \mathbf{0}, \ \mathbf{w}_2 \succeq \mathbf{0}, \end{aligned} \tag{2}$$

where $\mathbf{w}_1$ and $\mathbf{w}_2$ are the two pursued canonical projections; the constraints $\mathbf{w}_1 \succeq \mathbf{0}$, $\mathbf{w}_2 \succeq \mathbf{0}$ enforces that each element in $\mathbf{w}_1$ and $\mathbf{w}_2$ must be nonnegative. The constraints $\|\mathbf{w}_1\|^2 \le 1$, $\|\mathbf{w}_2\|^2 \le 1$ are the convex relaxations of the equality constraints $\|\mathbf{w}_1\|^2 = 1$, $\|\mathbf{w}_2\|^2 = 1$, which normalize the projections. P is a convex sparsity-inducing penalty with c_1 and c_2 being two sparsity controllers. Optimization with respect to $\mathbf{w}_1$ and $\mathbf{w}_2$ involves iteratively fixing one and solves the other.

3 Group-sparse Nonnegative supervised CCA (GNCCA)

GNCCA combines the benefits of SMVCCA and sparse nonnegative CCA by introducing a new group-sparse penalty, relaxing the normalization constraint, and, in particular, encouraging the relevance between features and class labels.

Following the notations in Eq. (1), let $\mathbf{Z} = [\mathbf{X}^T \ \mathbf{Y}^T]^T$ denote the stacking of the n K-View data samples $\mathbf{X}$ and their labels $\mathbf{Y}$. Correspondingly, let $\mathbf{C}_{xy}$ denote the stacking of the covariance matrix $\mathbf{C}$ and $\mathbf{Y}\mathbf{Y}^T$. In order to make GNCCA solvable after integrating group sparsity, the objective of SMVCCA, $trace(\mathbf{W}^T \mathbf{C}_{xy} \mathbf{W})$, is equivalently transformed to its Frobenius-Norm: $\|\mathbf{Z} - \mathbf{W}\mathbf{H}\|_F^2$, where $\mathbf{H}$ is the coefficient matrix, and $\mathbf{W}$ is the basis matrix. Based on this reformulation, GNCCA takes the form:

$$\begin{aligned} \min \ & \frac{1}{2}\|\mathbf{Z} - \mathbf{W}\mathbf{H}\|_F^2 + \alpha\|\mathbf{H}\|_F^2 + \beta \sum_{k=1}^{K} \|\mathbf{W}_{(k)}\|_{1,\infty}, \\ s.t. \ & \forall \|\mathbf{w}_i^{(k)}\|^2 \le 1 - \beta, k = 1, \cdots, K; i = 1, \cdots, r; \\ & \forall \|\mathbf{w}_i^{(y)}\|^2 = 1 - \beta, \ \ i = 1, \cdots, r; \\ & \mathbf{H} \ge \mathbf{0}, \ \mathbf{W} \ge \mathbf{0}, \end{aligned} \tag{3}$$

where $\mathbf{W} \in \mathcal{R}^{m \times r}$, $\mathbf{H} \in \mathcal{R}^{r \times n}$, parameters α, β control the relative influence of each penalty term, K is the number of feature views, r is the dimension of reduced representation, $r \ll m$, and $\|\|_{1,\infty}$ refers to $\ell_{1,\infty}$-norm. The nonnegative constraints $\mathbf{H} \ge \mathbf{0}$, $\mathbf{W} \ge \mathbf{0}$ ensure that both the canonical correlations of training and testing data are positive. The penalty $\|\mathbf{H}\|_F^2$ is to avoid an arbitrarily large $\mathbf{H}$.

The uniqueness of GNCCA lies in two aspects. First is the group-sparsity penalty on the K-View basis $\mathbf{W}_{(1:K)}$: $\sum_{k=1}^{K} \|\mathbf{W}_{(k)}\|_{1,\infty}$. Each $\ell_{1,\infty}$-norm is defined by: $\|\mathbf{W}_{(k)}\|_{1,\infty}$ $= \sum_{i=1}^{r} \|\mathbf{w}_i^{(k)}\|_\infty = \|\mathbf{w}_1^{(k)}\|_\infty + \cdots \|\mathbf{w}_r^{(k)}\|_\infty$, which is the sum of vector ℓ_∞-norms of its columns. Such $\ell_{1,\infty}$-norm is used to promote as many zero columns as possible in $\mathbf{W}_{(1)}, \cdots \mathbf{W}_{(K)}$, which indicates that only the correlations of the non-zero feature views are maximized. In this way, GNCCA captures both the sharing among modalities

and the uniqueness of each modality. $\|\mathbf{w}_i^{(k)}\|^2 \leq 1-\beta$ is to ensure that the correlations are normalized. In particular, we change the relaxation in Eq. (2), $\forall\|\mathbf{w}_i^{(k)}\|^2 \leq 1$, to be $\forall\|\mathbf{w}_i^{(k)}\|^2 \leq 1-\beta$. This is because the $\ell_{1,\infty}$ penalty is related to the constraints $\forall\|\mathbf{w}_i^{(k)}\|^2 \leq 1$. The insight is that for sparse features, the ℓ_2-norm of these features should be smaller than the norms of dense features.

Secondly, we keep the normalization for the label-view basis: $\forall\|\mathbf{w}_i^{(y)}\|^2 = 1-\beta$, which ensures that the selected features are closely correlated with the class labels.

Problem (3) is optimized by adapting the efficient vector-block coordinate descent (BCD) method [9], which updates one column of a factor matrix at each step while fixing all other values. The optimization randomly initializes $\mathbf{W}$ and $\mathbf{H}$, and then updates each vector of $\mathbf{H}$, $\mathbf{W}_{(1:K)}$ and $\mathbf{W}_{(y)}$ in three steps, until either the objective function is below a preset threshold or the maximum number of iterations has been reached.

Step 1: Fixing $\mathbf{W}$, update each row vector $\mathbf{h}_{i\cdot} \in \mathbb{R}^{1\times n}, i = 1, \cdots, r$ as

$$\mathbf{h}_{i\cdot} \leftarrow \arg\min_{\mathbf{h}\geq 0} \frac{1}{2}\|\mathbf{R}_i - \mathbf{w}_i\mathbf{h}\|_F^2 + \alpha\|\mathbf{h}\|_2^2, \tag{4}$$

where $\mathbf{R}_i = \mathbf{Z} - \sum_{j=1,j\neq i}^{r} \mathbf{w}_j\mathbf{h}_{j\cdot}$. This subproblem is solved in a closed form: $\mathbf{h}_{i\cdot} \leftarrow \left[\frac{\mathbf{w}_i^T\mathbf{R}_i}{2\alpha+\|\mathbf{w}_i\|^2}\right]_+$, where $[\,]_+$ denote the element-wise projection to nonnegative numbers.

Step 2: Fixing $\mathbf{H}$ and $\mathbf{W}_{(y)}$, update each column vector $\mathbf{w}_i^{(k)},\ k = 1, \cdots, K$ as

$$\mathbf{w}_i^{(k)} \leftarrow \arg\min_{\mathbf{w}\geq 0, \|\mathbf{w}\|^2\leq 1-\beta} \frac{1}{2}\|\mathbf{R}_i^{(k)} - \mathbf{w}^{(k)}\mathbf{h}_{i\cdot}\|_F^2 + \beta\|\mathbf{w}^{(k)}\|_\infty, \tag{5}$$

where $\mathbf{R}_i^{(k)} = \mathbf{Z}^{(k)} - \sum_{j=1,j\neq i}^{r} \mathbf{w}_j^{(k)}\mathbf{h}_{j\cdot}$. This subproblem can be efficiently solved by first finding the solution $\mathbf{w}_i^{(k)*}$ of the problem described in [10], then normalizing $\mathbf{w}_i^{(k)*}$ such that $\|\mathbf{w}_i^{(k)*}\|^2 \leq 1-\beta$, and finally updating $\mathbf{w}_i^{(k)*}$ as $\mathbf{w}_i^{(k)**} = \left[\frac{\mathbf{R}_i^{(k)}\mathbf{h}_{i\cdot}^T}{\|\mathbf{h}_{i\cdot}\|^2}\right]_+ - \mathbf{w}_i^{(k)*}$.

Step 3: Fixing $\mathbf{W}_{(1:K)}$ and $\mathbf{H}$, update $\mathbf{W}_{(y)}$ as

$$\mathbf{w}_i^{(y)} \leftarrow \arg\min_{\mathbf{w}\geq 0, \|\mathbf{w}\|^2=1-\beta} \frac{1}{2}\|\mathbf{R}_i^{(y)} - \mathbf{w}^{(y)}\mathbf{h}_{i\cdot}\|_F^2, \tag{6}$$

where $\mathbf{R}_i^{(y)} = \mathbf{Z}^{(y)} - \sum_{j=1,j\neq i}^{r} \mathbf{w}_j^{(y)}\mathbf{h}_{j\cdot}$. The solution of this subproblem is to first update $\mathbf{w}$ as $\mathbf{w}_i^{(y)} \leftarrow \left[\frac{\mathbf{R}_i^{(y)}\mathbf{h}_{i\cdot}^T}{\|\mathbf{h}_{i\cdot}\|^2}\right]_+$, and then normalize $\mathbf{w}_i^{(y)}$ by $\mathbf{w}_i^{(y)} \leftarrow \frac{\mathbf{w}_i^{(y)}}{\|\mathbf{w}_i^{(y)}\|}\sqrt{1-\beta}$.

The above optimization result of GNCCA leads to:

$$\mathbf{Z} = \begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix} \approx \begin{bmatrix}\mathbf{W}\\ \mathbf{W}_y\end{bmatrix}\mathbf{H},$$

which provides two approximate relationships: $\mathbf{X} \approx \mathbf{W}\mathbf{H}$ and $\mathbf{Y}^T \approx \mathbf{H}^T\mathbf{W}_y^T$. The former introduces a lower feature representation $\mathbf{H}$ in the reduced feature space $\mathbf{W}$ while the latter models how the lower representation $\mathbf{H}$ is regressed to class labels $\mathbf{Y}^T$. Given these two relationships, we rank the original jth feature according to the Variable Importance in the Projections (VIP) score [4], which determines the importance of original features based on their values in the projected subspace:

$$\pi_j = \sqrt{\frac{m\sum_{i=1}^{r}\|\mathbf{w}_y\|_2^i\mathbf{h}_{i\cdot}\mathbf{h}_{i\cdot}^T(\frac{w_{ji}}{\|\mathbf{w}_i\|})^2}{\sum_{i=1}^{r}\|\mathbf{w}_y\|_2^i\mathbf{h}_{i\cdot}\mathbf{h}_{i\cdot}^T}},$$

where $w_y^i \in \mathbf{W}_y$ is column entry of $\mathbf{W}_y$. To automatically determine the number of selected features, we define a threshold σ: $0 < \sigma < 1$. Only top features whose scores are no smaller than $\sigma \times \pi_{max}$ will be selected, where π_{max} is the largest VIP score. For simplicity, we call the GNCCA based feature selector GNCCA-VIP.

4 Experiments

4.1 Data Sets

To evaluate GNCCA-VIP, we chose three unique datasets that enabled us to address three of the most relevant problems in the disease domain of prostate cancer (CaP), i.e., predict biochemical recurrence, identifying CaP grade and localizing CaP. GNCCA-VIP is applied to select the most important ones from extracted imaging/non-imaging features such that the selected features can be used to better address the associated task of CaP prognosis.

We evaluate the proposed method by using D1 [2], D2 [2] and D3 [11] datasets (see Table 1). Representative images from D1 and D2 are shown in Fig. 1. D1 contains 21 CaP patients who experienced and 19 CaP patients who did not experience biochemical recurrence within 5 years of radical prostatectomy. For each patient, 242 histological features and 650 proteomic profiling values are extracted. D2 contains 33 MRI slices from 16 CaP patients. 22 slices have prostate cancer with a Gleason grade [12] of 3, while the other 11 have a grade of 4. On each slice, 112 textural features and 38 kinetic features are extracted from T2w and DCE MRI, respectively. D3 contains 2901 CaP/non-CaP voxels from 36 patients that were annotated by an expert clinician. From within each voxel, 6 MRS features and 58 MRI features are extracted.

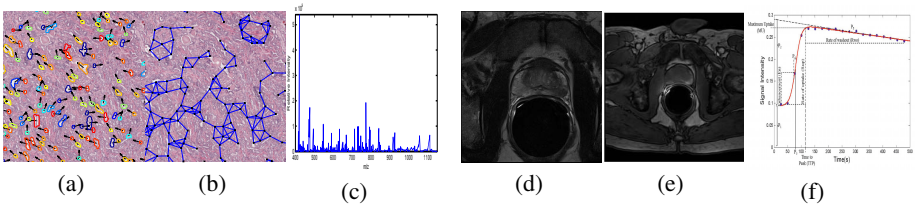
Table 1. Summary of the benchmark data sets

Data set	#Views	#Features	#Patients	Task
D1 [2]	2	892(242+650)	40	predict biochemical recurrence
D2 [2]	2	150(112+38)	16	identify CaP grade
D3 [11]	2	64(6+58)	36	localize CaP

4.2 Experimental Settings

A Leave-M-Patients-Out cross validation scheme was used in conjunction with a linear Support Vector Machine (SVM) classifier [13] to evaluate the feature selection performance of GNCCA-VIP with the following comparative strategies: i) All-features, ii) FisherScore [5], iii) T-test [6], iv) mRMR [7], v) PCA-VIP [4], and vi) SMVCCA-VIP [2]. The parameter c of SVM is set to be $c = 1$. The critical parameters of GNCCA, including β, r and σ, are determined via Leave-1-Patient-out cross validation. The controller of $\|\mathbf{H}\|_F^2$, α, is fixed at 0.1. The number of patients left out, M, varies according to the number of total patients in each dataset.

Each experiment was repeated 50 times, and the average classification accuracy as well as variations were reported. All the experiments were conducted within a Matlab environment on a 64-bit Linux machine with 4-core CPU and 4G memory.

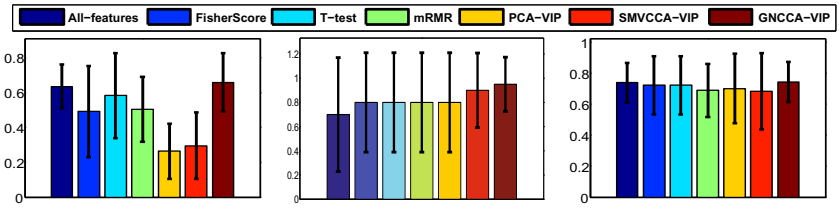


(a) (b) (c) (d) (e) (f)

Fig. 1. D1 histology dataset: (a) textural features extracted on excised gland orientations, (b) structural features, and (c) Proteomic profiling values. D2 dataset: (d) T2w MRI, (e) DCE MRI, and (f) Kinetic curve extracted from DCE MRI.

4.3 Experiment 1: GNCCA-VIP versus Other Feature Selection Schemes

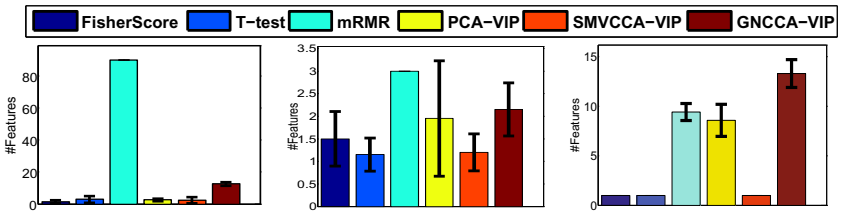
Fig. 2 illustrates the classification results obtained with the compared feature selection approaches in conjunction with linear SVM. Fig. 3 shows the corresponding number of selected features with each selection approach. On D1, GNCCA-VIP achieves better classification than All-features while selecting only 2% of the features (18 out of 892). Although other methods (except mRMR) select fewer features, their classification results are obviously worse. On D2, GNCCA-VIP obtains the highest classification results with only selecting an average of 2 features from 150 features. On D3, GNCCA-VIP selects an average of 13 features (out of the 64 features) and achieves the same classification result as using all features.



(a) D1 (b) D2 (c) D3

Fig. 2. Comparing the prognosis accuracies (and standard deviations (STD)) with the features selected via All-features, FisherScore, T-test, mRMR, PCA-VIP, SMVCCA-VIP and GNCCA-VIP on the three dataset. The same linear SVM is used as the classifier here. The proposed GNCCA-VIP approach achieves highest accuracies (with lowest STDs) in all the prognosis tasks.

The results also reveal that GNCCA-VIP is likely to select a comparable number of features from each modality. For example, on D3, GNCCA-VIP is most likely to select all 6 MRS features and then select 6 or 7 MRI features that are related to the 6 MRS features. By contrast, the other feature selection methods are likely to only select MRI features, which is clearly a bias towards the dominant MRI features. All these results demonstrate the ability of GNCCA-VIP in identifying features that are discriminative of provided class labels and leads to better view association.

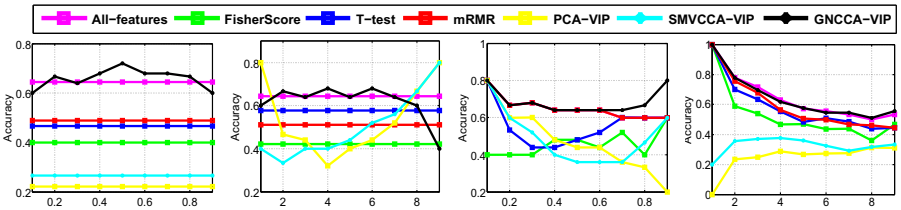


(a) D1 (b) D2 (c) D3

Fig. 3. The average number of features (STD) adaptively selected via FisherScore, T-test, mRMR, PCA-VIP, SMVCCA-VIP and GNCCA-VIP on achieving the reported prognosis accuracies in Fig. 2. Note that at each time mRMR selects the same number of features on D1 and D2.

4.4 Experiment 2: Parameter Sensitivity of GNCCA-VIP

Fig. 4 shows the sensitivity of GNCCA-VIP to four critical parameters that must be manually tuned: i) the group-sparsity controller β in Eq. (3), ii) the intrinsic dimensionality r in Eq. (3), iii) the threshold σ that controls the number of selected features, and iv) M, the number of patients left out. The D1 dataset is used for this experiment. As we can see, GNCCA-VIP is relatively insensitive to β, r and σ. As M increases, indicating that more patients are being used as testing samples and fewer subjects are being retained for training purposes, all the classification results gradually degrade. However, GNCCA-VIP consistently leads to favourable classification result compared to using all features.



(a) varying β (b) varying r (c) varying σ (d) varying M

Fig. 4. Testing the influence of parameters β , r , σ and M. **(a)** varying β from 0.1 to 0.9 by fixing $r = 1, \sigma = 0.9, M = 1$. **(b)** varying r from 1 to 9 by fixing $\beta = 0.7, \sigma = 0.9, M = 5$. **(c)** varying σ from 0.1 to 0.9 by fixing $\beta = 0.7, r = 2, M = 1$. **(d)** varying M from 1 to 9 by fixing $r = 1, \sigma = 0.7, \beta = 0.9$. Note that the D1 dataset is used for this experiment.

5 Concluding Remarks

This paper presents GNCCA-VIP, a novel method for multimodal feature selection that incorporates nonnegativity and group-sparse constraints into supervised CCA. The nonnegativity ensures that latent components are always positively correlated. Group sparsity incorporates strengths from both between- and within- view sparsity, thus allowing

for simultaneous view association and single-view feature selection. In our experiments, the features selected via GNCCA-VIP result in improved classification performance in CaP prognosis prediction compared to using either the entire feature set or using other state-of-the-art feature selection schemes. Hence, this approach may be a better choice in the context of combining multi-scale, multi-modal data.

References

1. Kettenring, J.R.: Canonical analysis of several sets of variables. Biometrika 58(3), 433–451 (1971)
2. Singanamalli, A., Wang, H., et al.: Supervised multi-view canonical correlation analysis: Fused multimodal prediction of disease diagnosis and prognosis. In: SPIE Medical Imaging, vol. 9038 (2014)
3. Witten, D.M., Tibshirani, R.J.: Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical Applications in Genetics and Molecular Biology 8(1), 1–27 (2009)
4. Ginsburg, S., Tiwari, P., Kurhanewicz, J., Madabhushi, A.: Variable ranking with pca: Finding multiparametric mr imaging markers for prostate cancer diagnosis and grading. In: Madabhushi, A., Dowling, J., Huisman, H., Barratt, D. (eds.) Prostate Cancer Imaging 2011. LNCS, vol. 6963, pp. 146–157. Springer, Heidelberg (2011)
5. Duda, R.O., Hart, P., Stork, D.: Pattern classification. Journal of Classification 24(2), 305–307 (2007)
6. Student: The problem error of a mean. Biometrika 6, 1–25 (1908)
7. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI) 27(8), 1226–1238 (2005)
8. Ye, J., Liu, J.: Sparse methods for biomedical data. SIGKDD 14(1), 4–15 (2012)
9. Jingu Kim, R.M., Park, H.: Group sparsity in nonnegative matrix factorization. In: SIAM International Conference on Data Mining (SDM), pp. 851–862 (2012)
10. Duchi, J., et al.: Efficient projections onto the l1-ball for learning in high dimensions. In: The 25th International Conference on Machine Learning (ICML), pp. 272–279 (2008)
11. Tiwari, P., et al.: Multimodal wavelet embedding representation for data combination (maweric): integrating magnetic resonance imaging and spectroscopy for prostate cancer detection. NMR in Biomedicine 25(4), 607–619 (2012)
12. Gleason, D.F.: The veteran's administration cooperative urologic research group: histologic grading and clinical staging of prostatic carcinoma. In: Urologic Pathology: The Prostate, pp. 171–198 (1977)
13. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)

Clustering-Induced Multi-task Learning for AD/MCI Classification

Heung-Il Suk and Dinggang Shen

Biomedical Research Imaging Center, University of North Carolina at Chapel Hil
{hsuk,dgshen}@med.unc.edu

Abstract. In this work, we formulate a clustering-induced multi-task learning method for feature selection in Alzheimer's Disease (AD) or Mild Cognitive Impairment (MCI) diagnosis. Unlike the previous methods that often assumed a unimodal data distribution, we take into account the underlying multipeak[1] distribution of classes. The rationale for our approach is that it is likely for neuroimaging data to have multiple peaks or modes in distribution due to the inter-subject variability. In this regard, we use a clustering method to discover the multipeak distributional characteristics and define subclasses based on the clustering results, in which each cluster covers a peak. We then encode the respective subclasses, *i.e.*, clusters, with their unique codes by imposing the subclasses of the same original class close to each other and those of different original classes distinct from each other. We finally formulate a multi-task learning problem in an $\ell_{2,1}$-penalized regression framework by taking the codes as new label vectors of our training samples, through which we select features for classification. In our experimental results on the ADNI dataset, we validated the effectiveness of the proposed method by achieving the maximal classification accuracies of 95.18% (AD/Normal Control: NC), 79.52% (MCI/NC), and 72.02% (MCI converter/MCI non-converter), outperforming the competing single-task learning method.

1 Introduction

From a computational modeling perspective, while the feature dimension of neuroimaging data is high in nature, we have a very limited number of observations/samples available. This so-called "small-*n*-large-*p*" problem has been of a great challenge in the field to build a robust model that can correctly identify a clinical label of a subject, *e.g.*, AD, MCI, Normal Control (NC) [10]. For this reason, reducing the feature dimensionality, by which we can mitigate the overfitting problem and improve a model's generalizability, has been considered as a prevalent step in building a computer-aided AD diagnosis system as well as neuroimaging analysis [6]. On the other hand, pathologically, since the disease-related atrophy or hypo-metabolism could happen in the part of a Region Of Interest (ROI), or cover small regions of multiple ROIs, it is difficult to predefine

[1] Even though the term of "multimodal distribution" is generally used in the literature, in order to avoid the confusion with the "multimodal" neuroimaging, we use the term of "multipeak distribution" throughout the paper.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 393–400, 2014.

ROIs, and thus important to consider the whole brain features and then select the most informative ones for better diagnosis.

The main limitation of the previous methods of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), and an embedded method such as ℓ_1-penalized regression model is that they consider a single mapping or a single weight coefficient vector in reducing the dimensionality. But, if the underlying data distribution is not unimodal, *e.g.*, mixture of Gaussians, then these methods would fail to find the proper mapping or weighting functions, and thus result in performance degradation. In this regard, Zhu and Martinez proposed a Subclass Discriminant Analysis (SDA) [12] that first clustered samples of each class and then reformulated the conventional LDA by regarding clusters as subclasses. Recently, Liao *et al.* applied the SDA method to segment prostate MR images and showed the effectiveness of the subclass-based approach [5].

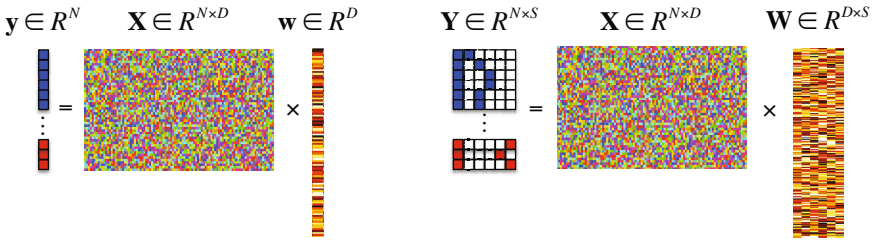
In this paper, we propose a novel method of feature selection for AD/MCI diagnosis by integrating the embedded method with the subclass-based approach. The motivation of clustering samples per class is the potential heterogeneity within a group, which may result from (1) a wrong clinical diagnosis; (2) different sub-types in AD (e.g., amnestic/non-amnestic); (3) conversion of MCI non-converter or NC to AD after the follow-up time. Specifically, we first divide each class into multiple subclasses by means of clustering, with which we can approximate the inherent multipeak data distribution of a class. Note that we regard each cluster as a subclass by following Zhu and Martinez's work [12]. Based on the clustering results, we encode the respective subclasses with their unique codes, for which we impose the subclasses of the same original class close to each other and those of different original classes distinct from each other. By setting the codes as new labels of our training samples, we finally formulate a multi-task learning problem in an $\ell_{2,1}$-penalized regression framework that takes into account the multipeak data distributions, and thus help enhance the diagnostic performances.

2 Materials and Image Processing

We use the ADNI dataset publicly available on the web[2]. Specifically, we consider only the baseline Magnetic Resonance Imaging (MRI) and 18-Fluoro-DeoxyGlucose (FDG) Positron Emission Tomography (PET) data acquired from 51 AD, 99 MCI, and 52 NC subjects. For the MCI subjects, they were further clinically subdivided into 43 MCI Converters (MCI-C) and 56 MCI Non-Converters (MCI-NC), who progressed and did not progress to AD in 18 months, respectively.

The MR images were preprocessed by applying the prevalent procedures of Anterior Commissure (AC)-Posterior Commissure (PC) correction, skull-stripping, and cerebellum removal. Specifically, we used MIPAV software[3] for AC-PC correction, resampled images to 256×256×256, and applied N3 algorithm [8] for intensity inhomogeneity correction. Then, structural MR images were segmented into three tissue types of Gray Matter (GM), White Matter (WM) and CSF with FAST in FSL package[4]. We finally parcellated them into 93 ROIs by warping Kabani *et al.*'s atlas [4] to each subject's

[2] Available at '`http://www.loni.ucla.edu/ADNI`'
[3] Available at '`http://mipav.cit.nih.gov/clickwrap.php`'
[4] Available at '`http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/`'



(a) Single-task learning (b) Multi-task learning

Fig. 1. In the response vector/matrix, the colors of blue, red, and white represent 1, -1, and 0, respectively. In multi-task learning, each row of the response matrix represents a newly defined sparse code for each sample by the proposed method.

brain space. Regarding FDG-PET images, they were rigidly aligned to the respective MR images, and then applied parcellation propagated from the atlas by registration. For each ROI, we used the GM[5] tissue volume from MRI, and the mean intensity from FDG-PET as features. Therefore, we have 93 features from an MR image and the same dimensional features from an FDG-PET image.

3 Method

Throughout the paper, we denote matrices as boldface uppercase letters, vectors as boldface lowercase letters, and scalars as normal italic letters, respectively. For a matrix $\mathbf{X} = [x_{ij}]$, its *i*-th row and *j*-th column are denoted as $\mathbf{x}^i$ and $\mathbf{x}_j$, respectively. We further denote the Frobenius norm and $\ell_{2,1}$-norm of a matrix $\mathbf{X}$ as $\|\mathbf{X}\|_F = \sqrt{\sum_i \|\mathbf{x}^i\|_2^2} = \sqrt{\sum_j \|\mathbf{x}_j\|_2^2}$ and $\|\mathbf{X}\|_{2,1} = \sum_i \|\mathbf{x}^i\|_2 = \sum_i \sqrt{\sum_j x_{ij}^2}$, respectively, and the ℓ_1-norm of a vector as $\|\mathbf{w}\|_1 = \sum_i |w_i|$.

3.1 Preliminaries

Let $\mathbf{X} \in R^{N \times D}$ and $\mathbf{y} \in R^N$ denote, respectively, the D neuroimaging features and clinical labels of N samples. Assuming that the clinical label can be represented by a linear combination of the neuroimaging features, many research groups have utilized a least square regression model with various regularization terms. In particular, despite its simple form, the ℓ_1-penalized linear regression model has been widely and successfully used in the literature [1, 11], as formulated as follows:

$$\min_{\mathbf{w}} \ \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda_1 \|\mathbf{w}\|_1 \tag{1}$$

[5] Based on the previous studies that showed the relatively high relatedness of GM compared to WM and CSF, we use only features from GM in classification.

where λ_1 denotes a sparsity control parameter. Since this method finds a single optimal weight coefficient vector $\mathbf{w}$ that regresses the target response vector $\mathbf{y}$, it is classified into a single-task learning (Fig. 1(a)) in machine learning.

If there exists additional class-related information, then we can further extend the ℓ_1-penalized linear regression model into a more sophisticated $\ell_{2,1}$-penalized one as follows:

$$\min_{\mathbf{W}} \|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda_2 \|\mathbf{W}\|_{2,1} \tag{2}$$

where $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_S] \in R^{N \times S}$ is a target response matrix, $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_S] \in R^{D \times S}$ is a weight coefficient matrix, S is the number of response variables, and λ_2 denotes a group sparsity control parameter. In machine learning, this framework is classified into a multi-task learning[6] (Fig. 1(b)) because it needs to find a set of weight coefficient vectors by regressing multiple response values, simultaneously.

3.2 Clustering-Induced Multi-task Learning

Because of the inter-subject variability [3, 7], it is likely for neuroimaging data to have multiple peaks in distribution. In this paper, we argue that it is necessary to consider the underlying multipeak data distribution in feature selection. To this end, we propose to divide classes into subclasses and to utilize the resulting subclass information for guiding feature selection by means of a multi-task learning.

To divide the training samples of each original class into their respective subclasses, we exploit a clustering technique. Specifically, thanks to its simplicity and computational efficiency, especially in a high dimensional space, we use a K-means algorithm. Note that the resulting clusters are regarded as subclasses, following Zhu and Martinez's work [12]. We then encode the subclasses with their unique labels, for which we use *discriminative* sparse codes to enhance classification performance. Let $K_{(+)}$ and $K_{(-)}$ denote, respectively, the number of clusters/subclasses for the original classes of '$+$' and '$-$'. Without loss of generality, we define sparse codes for the subclasses of the original classes of '$+$' and '$-$' as follows:

$$\mathbf{s}_l^{(+)} = \begin{bmatrix} +1 & \mathbf{z}_l^{(+)} & \mathbf{0}_{K_{(-)}} \end{bmatrix} \tag{3}$$

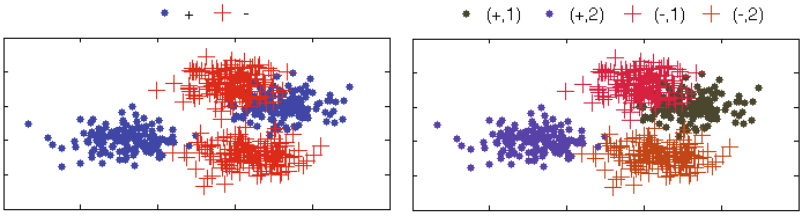
$$\mathbf{s}_m^{(-)} = \begin{bmatrix} -1 & \mathbf{0}_{K_{(+)}} & \mathbf{z}_m^{(-)} \end{bmatrix} \tag{4}$$

where $l \in \{1, \cdots, K_{(+)}\}$, $m \in \{1, \cdots, K_{(-)}\}$, $\mathbf{0}_{K_{(+)}}$ and $\mathbf{0}_{K_{(-)}}$ denote, respectively, zero row vectors with $K_{(+)}$ and $K_{(-)}$ elements, and $\mathbf{z}_l^{(+)} \in \{0, 1\}^{K_{(+)}}$ and $\mathbf{z}_m^{(-)} \in \{0, -1\}^{K_{(-)}}$ denote, respectively, indicator row vectors in which only the l-th/m-th element is set to 1/-1 and the others are 0. Thus, the full code set is defined as follows:

$$\mathbb{S} = \{\mathbf{s}_1^{(+)}, \cdots, \mathbf{s}_l^{(+)}, \cdots, \mathbf{s}_{K_{(+)}}^{(+)}, \mathbf{s}_1^{(-)}, \cdots, \mathbf{s}_m^{(-)}, \cdots, \mathbf{s}_{K_{(-)}}^{(-)}\}. \tag{5}$$

Fig. 2 presents a simple toy example of finding subclasses and defining the respective sparse code vectors. It is noteworthy that in our sparse code set, we reflect the original

[6] To regress each response value is considered as a task.



(a) Data distribution of '+'/'-' classes (b) Subclasses (original class, cluster #)

Fig. 2. A toy example of finding subclasses and defining the respective sparse code vectors. $(+,1) : \mathbf{s}_1^{(+)} = \begin{bmatrix}+1 & +1 & 0 & 0 & 0\end{bmatrix}$, $(+,2) : \mathbf{s}_2^{(+)} = \begin{bmatrix}+1 & 0 & +1 & 0 & 0\end{bmatrix}$, $(-,1) : \mathbf{s}_1^{(-)} = \begin{bmatrix}-1 & 0 & 0 & -1 & 0\end{bmatrix}$, and $(-,2) : \mathbf{s}_2^{(-)} = \begin{bmatrix}-1 & 0 & 0 & 0 & -1\end{bmatrix}$.

label information to our new codes by setting the first element of the sparse codes with their original label. Furthermore, by setting the indicator vectors $\{\mathbf{z}_l^{(+)}\}_{l=1}^{K_{(+)}}$ and $\{\mathbf{z}_m^{(-)}\}_{m=1}^{K_{(-)}}$ to be positive and negative, respectively, the distances become close among the subclasses of the same original class while distant among the subclasses of the different original classes.

Using the newly defined sparse codes, we assign a new label vector $\mathbf{y}^i$ to a training sample $\mathbf{x}^i$ as follows:

$$\mathbf{y}^i = \mathbf{s}_{\gamma_i}^{(y_i)} \tag{6}$$

where $y_i \in \{+,-\}$ is the original label of the training sample $\mathbf{x}^i$, and γ_i denotes the cluster to which the sample $\mathbf{x}^i$ was assigned by the K-means algorithm. In this way, we extend the original scalar labels of +1 or -1 into sparse code vectors in $\mathbb{S}$.

Thanks to our new sparse codes, it becomes natural to convert a single-task learning in Eq. (1) into a multi-task learning in Eq. (2) by replacing the original label vector $\mathbf{y}$ in Eq. (1) with a matrix $\mathbf{Y} = \left[\mathbf{y}^i\right]_{i=1}^N \in \{-1,0,1\}^{N\times(1+K_{(+)}+K_{(-)})}$. Therefore, we have now $(1+K_{(+)}+K_{(-)})$ tasks. Note that the task of regressing the first column response vector $\mathbf{y}_1$ corresponds to our binary classification problem between the original classes of '+' and '−'. Meanwhile, the tasks of regressing the remaining column vectors $\{\mathbf{y}_i\}_{i=2}^{1+K_{(+)}+K_{(-)}}$ formulate new binary classification problems between one subclass and all the other subclasses. It should be noted that unlike the single-task learning that finds a single mapping $\mathbf{w}$ between regressors $\mathbf{X}$ and the response $\mathbf{y}$, the clustering-induced multi-task learning finds multiple mappings $\{\mathbf{w}_1,\cdots,\mathbf{w}_{(1+K_{(+)}+K_{(-)})}\}$, and thus allows us to efficiently use the underlying multipeak data distribution in feature selection.

3.3 Feature Selection and Classifier Learning

Because of the $\ell_{2,1}$-norm regularizer in our objective function of Eq. (2), after finding the optimal solution, we have some zero row-vectors in $\mathbf{W}$. In terms of the linear regression, the corresponding features are not informative in regressing the response

Table 1. A summary of the performances for AD/NC classification

Method	Modality	ACC (%)	SEN (%)	SPEC (%)	BAC (%)	AUC (%)
STL	MRI	90.45±6.08	82.67	**98.33**	90.50	93.55
	PET	86.27±8.59	82.00	90.33	86.17	90.12
	MRI+PET	92.27±5.93	90.00	94.67	92.33	94.91
CIMTL	MRI	93.27±6.33	88.33	**98.33**	93.33	94.19
	PET	89.27±7.43	90.00	88.33	89.17	91.67
	MRI+PET	**95.18±6.65**	**94.00**	96.33	**95.17**	**96.15**

values. In this regard, we finally select the features whose weight coefficient vector is non-zero, *i.e.*, $\|\mathbf{w}^i\|_2 > 0$. With the selected features, we then train a linear Support Vector Machine (SVM) for making a diagnostic decision.

4 Experimental Results and Analysis

4.1 Experimental Setting

We considered three binary classification problems: AD/NC, MCI/NC, and MCI-C/MCI-NC. In the classification of MCI/NC, we labeled both MCI-C and MCI-NC as MCI. Due to the limited number of samples, we applied a 10-fold cross-validation technique in each binary classification problem. Specifically, we randomly partitioned the samples of each class into 10 subsets with approximately equal size without replacement. We then used 9 out of 10 subsets for training and the remaining one for testing. For performance comparison, we took the average of the 10 cross-validation results.

Regarding model selection, *i.e.*, number of clusters K, sparsity control parameters of λ_1 in Eq. (1) and λ_2 in Eq. (2), and the soft margin parameter C in SVM [2] , we further split the training samples into 5 subsets for nested cross-validation. To be more specific, we defined the spaces of the model parameters as follows: $K \in \{1, 2, 3, 4, 5\}$, $C \in \{2^{-10}, \ldots, 2^5\}$, $\lambda_1 \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.5\}$, and $\lambda_2 \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.5\}$. The parameters that achieved the best classification accuracy in the inner cross-validation were finally used in testing.

To validate the effectiveness of the proposed Clustering-Induced Multi-Task Learning (CIMTL) method, we compared it with the Single-Task Learning (STL) method that used only the original class label as the target response vector. For each set of experiments, we used 93 MRI features and/or 93 PET features as regressors in the respective least square regression models. Regarding the neuroimaging fusion of MRI and PET [9], we constructed a long feature vector by concatenating features of the modalities. It should be noted that the only difference between the proposed CIMTL method and the competing STL method lies in the way of selecting features, *i.e.*, single-task learning vs. multi-task learning. We used five quantitative metrics for comparison: ACCuracy (ACC), SENsitivity (SEN), SPECificity (SPEC), Balanced ACcuracy (BAC), and Area Under the receiver operating characteristic Curve (AUC).

Table 2. A summary of the performances for MCI/NC classification

Method	Modality	ACC (%)	SEN (%)	SPEC (%)	BAC (%)	AUC (%)
STL	MRI	74.85±5.92	80.67	64.00	72.33	76.55
	PET	69.51±10.11	74.78	59.67	67.22	73.54
	MRI+PET	74.85±3.91	84.78	56.00	70.39	78.79
CIMTL	MRI	76.82±7.15	85.78	59.67	72.72	77.84
	PET	74.18±7.18	81.89	59.67	70.78	72.73
	MRI+PET	**79.52±5.39**	**88.89**	**62.00**	**75.44**	**77.91**

Table 3. A summary of the performances for MCI-C/MCI-NC classification

Method	Modality	ACC (%)	SEN (%)	SPEC (%)	BAC (%)	AUC (%)
STL	MRI	56.98±20.61	51.00	60.67	55.83	58.85
	PET	61.58±17.79	55.00	66.00	60.50	60.63
	MRI+PET	64.62±14.04	**62.50**	66.00	64.25	63.87
CIMTL	MRI	61.60±13.12	44.00	75.67	59.83	60.76
	PET	66.73±11.32	39.00	**88.00**	63.50	65.57
	MRI+PET	**72.02±13.80**	58.00	82.67	**70.33**	**69.64**

4.2 Classification Results and Discussion

We summarized the performances of the competing methods with various modalities for AD and NC classification in Table 1. The proposed method showed the mean ACCs of 93.27% (MRI), 89.27% (PET), and 95.18% (MRI+PET). Compared to the STL method that showed the ACCs of 90.45% (MRI), 86.27% (PET), and 92.27% (MRI+PET), the proposed CIMTL method improved by 2.82% (MRI), 3% (PET), and 2.91% (MRI+PET). The proposed CIMTL method achieved higher AUC values than the STL method for all the cases. It is also remarkable that, except for the metric of SPEC with PET, 90.33% (STL) vs. 88.33% (CIMTL), the proposed CIMTL method consistently outperformed the competing STL method over all the metrics and modalities.

In the discrimination of MCI from NC, as reported in Table 2, the proposed CIMTL method showed the ACCs of 76.82% (MRI), 74.18% (PET), and 79.52% (MRI+PET). Meanwhile, the STL method showed the ACCs of 74.85% (MRI), 69.51% (PET), and 74.85% (MRI+PET). Again, the proposed CIMTL method outperformed the STL method by improving ACCs of 1.97% (MRI), 4.67% (PET), and 4.67% (MRI+PET), respectively. We believe that the high sensitivities and the low specificities for both competing methods resulted from the imbalanced data between MCI and NC. In the metrics of BAC and AUC that somehow reflect the imbalance of the test samples, the proposed method achieved the best BAC of 75.44% and the best AUC of 77.91% with MRI+PET.

Lastly, we conducted experiments of MCI-C and MCI-NC classification, and compared the results in Table 3. The proposed CIMTL method achieved the best ACC of 72.02%, the best BAC of 70.33%, and the best AUC of 69.64% with MRI+PET. In line with the fact that the classification between MCI-C and MCI-NC is the most important for early diagnosis and treatment, it is remarkable that compared to the STL method, the propose method improved the ACCs by 4.62% (MRI), 5.15% (PET), and 7.4% (MRI+PET), respectively.

For interpretation of the selected features, we built a histogram of the frequency of the selected ROIs of MRI and PET over CVs per binary classification. By setting

the mean frequency as the threshold, features from the following ROIs were mostly selected: subcortical regions (*e.g.*, amygdala, hippocampus, parahippocampal gyrus) and temporal lobules (*e.g.*, superior/middle temporal gyrus, temporal pole).

Regarding the identified subclasses, we computed the statistics (mean±std) of the optimal number of clusters determined in our cross-validation: 2.5±1.7/2.5±1.2 (AD/NC), 3.1±1.1/2.9±1.2 (MCI/NC), 3.4±0.8/3.8±1.3 (MCI-C/MCI-NC). Based on these statistics, we can say that there exists heterogeneity in a group, and by reflecting such information in feature selection, we could improve the diagnostic accuracy.

5 Conclusion

In this paper, we proposed a novel method that formulates a clustering-induced multi-task learning by taking into account the underlying multipeak data distribution of the original classes. In our experiments on the ADNI dataset, we proved the validity of the proposed method and showed its significantly better performance than the competing methods in the three binary classifications of AD/NC, MCI/NC, and MCI-C/MCI-NC.

References

1. de Brecht, M., Yamagishi, N.: Combining sparseness and smoothness improves classification accuracy and interpretability. NeuroImage 60(2), 1550–1561 (2012)
2. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2(2), 121–167 (1998)
3. Fotenos, A., Snyder, A., Girton, L., Morris, J., Buckner, R.: Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. Neurology, 1032–1039 (2005)
4. Kabani, N., MacDonald, D., Holmes, C., Evans, A.: A 3D atlas of the human brain. NeuroImage 7(4), S717 (1998)
5. Liao, S., Gao, Y., Shi, Y., Yousuf, A., Karademir, I., Oto, A., Shen, D.: Automatic prostate mr image segmentation with sparse label propagation and domain-specific manifold regularization. In: Gee, J.C., Joshi, S., Pohl, K.M., Wells, W.M., Zöllei, L. (eds.) IPMI 2013. LNCS, vol. 7917, pp. 511–523. Springer, Heidelberg (2013)
6. Mwangi, B., Tian, T., Soares, J.: A review of feature reduction techniques in neuroimaging. Neuroinformatics, 1–16 (2013)
7. Noppeney, U., Penny, W.D., Price, C.J., Flandin, G., Friston, K.J.: Identification of degenerate neuronal systems based on intersubject variability. NeuroImage 30(3), 885–890 (2006)
8. Sled, J.G., Zijdenbos, A.P., Evans, A.C.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Transactions on Medical Imaging 17(1), 87–97 (1998)
9. Suk, H.I., Lee, S.W., Shen, D.: Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Structure and Function, 1–19 (2013)
10. Suk, H.I., Wee, C.Y., Shen, D.: Discriminative group sparse representation for mild cognitive impairment classification. In: Wu, G., Zhang, D., Shen, D., Yan, P., Suzuki, K., Wang, F. (eds.) MLMI 2013. LNCS, vol. 8184, pp. 131–138. Springer, Heidelberg (2013)
11. Varoquaux, G., Gramfort, A., Poline, J.B., Thirion, B.: Brain covariance selection: better individual functional connectivity models using population prior. In: Lafferty, J., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (eds.) Advanced in Neural Information Processing Systems, pp. 2334–2342 (2010)
12. Zhu, M., Martinez, A.M.: Subclass discriminant analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(8), 1274–1286 (2006)

A Novel Multi-relation Regularization Method for Regression and Classification in AD Diagnosis

Xiaofeng Zhu, Heung-Il Suk, and Dinggang Shen*

Department of Radiology and Biomedical Research Imaging Center,
University of North Carolina at Chapel Hill, USA
dgshen@med.unc.edu

Abstract. In this paper, we consider the joint regression and classification in Alzheimer's disease diagnosis and propose a novel multi-relation regularization method that exploits the relational information inherent in the observations and then combines it with an $\ell_{2,1}$-norm within a least square regression framework for feature selection. Specifically, we use three kinds of relationships: feature-feature relation, response-response relation, and sample-sample relation. By imposing these three relational characteristics along with the $\ell_{2,1}$-norm on the weight coefficients, we formulate a new objective function. After feature selection based on the optimal weight coefficients, we train two support vector regression models to predict the clinical scores of Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog) and Mini-Mental State Examination (MMSE), respectively, and a support vector classification model to identify the clinical label. We conducted clinical score prediction and disease status identification jointly on the Alzheimer's Disease Neuroimaging Initiative dataset. The experimental results showed that the proposed regularization method outperforms the state-of-the-art methods, in the metrics of correlation coefficient and root mean squared error in regression and classification accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve in classification.

Keywords: Alzheimer's disease, feature selection, sparse coding, manifold learning, MCI conversion.

1 Introduction

For the computer-aided Alzheimer's Disease (AD) or Mild Cognitive Impairment (MCI) diagnosis, the available sample size is usually small, but the feature dimension is high. For example, the sample size used in [7,21] was less than one hundred, while the feature dimension (including both Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) features) was hundreds or even thousands. The small sample size makes it difficult to build an effective model, and the high dimensionality of data leads to an overfitting problem. To this end, researchers mostly predefined the disease-related features and used such low-dimensional features in clinical label identification or clinical score prediction.

In the meantime, recent studies have shown that the feature selection helps overcome both problems of high dimensionality and small sample size, by removing uninformative features [14,16,13,19,20,18]. Moreover, among various feature selection

* Corresponding author.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 401–408, 2014.

techniques, manifold learning has been successfully used in either regression or classification [9,13,12,17]. For example, Cho *et al.* adopted a manifold harmonic transformation method on the cortical thickness data and achieved a sensitivity of 63% and a specificity of 76% on the dataset with 72 MCI Converters (MCI-C) and 131 MCI Non-Converters (MCI-NC) [3]. While most of the previous studies focused on identifying brain disease and estimating clinical scores separately [4], there have been also efforts to select joint features that could be used for both tasks simultaneously. For example, Zhang and Shen proposed a multi-task sparse feature selection method for joint disease status identification and clinical scores prediction, and showed that such combination can achieve better performance than performing them separately [15,21].

In line with Zhang and Shen's work, in this paper, we consider the prediction of both clinical scores and disease status jointly in a unified framework, as in [7,9]. However, unlike the previous manifold-based feature selection methods that considered only the manifold of the samples, but not manifold of either the features or the response variables, we propose a novel multi-relation regularization method. Specifically, we use the relational information inherent in the observations and combine it with an $\ell_{2,1}$-norm within a least square regression framework. The rationale for the proposed multi-relation regularization method is as follows: (1) If some features are related to each other, then the same or similar relation is expected to be preserved between the respective weight coefficients in a least square regression model. (2) Due to the algebraic operation in least square regression, *i.e.,* matrix multiplication, the weight coefficients are linked to the response variables via regressors, *i.e.,* feature vectors in our work. Therefore, it is natural to impose the relation between a pair of weight coefficients to be similar to the relation between the corresponding pair of target response variables. (3) As considered in many manifold learning methods [1,6,17], if a pair of samples are similar to each other, then their respective response values should be also similar to each other. By imposing these three relational characteristics along with the $\ell_{2,1}$-norm on the weight coefficients, we formulate a new objective function. We then select features to build classification and regression models for clinical label identification and clinical scores (Alzheimer's Disease Assessment Scale-Cognitive subscale: ADAS-Cog, Mini-Mental State Examination: MMSE) prediction, respectively.

2 Method

By taking the features as regressors and the concatenation of clinical scores (*e.g.,* ADAS-Cog, MMSE) and a class label as responses, we apply the proposed method to select features that are jointly used to represent clinical scores and class labels. Based on the selected features, we finally build clinical scores regression models and a clinical label identification model with Support Vector Regression (SVR) and Support Vector Classification (SVC), respectively.

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{Y} \in \mathbb{R}^{n \times c}$ denote d neuroimaging features and c clinical response values of n subjects or samples, respectively. In this work, we assume that the response values of clinical scores and a clinical label can be represented by a linear

combination of the features. Then, the problems of regressing clinical scores and identifying a class label can be formulated by a least square regression model as follows:

$$\mathcal{L}(\mathbf{W}) = \|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{c}(y_{ij} - \hat{y}_{ij})^2 \tag{1}$$

where $\|\cdot\|_F$ denotes a Frobenius norm, $\mathbf{W} \in \mathbb{R}^{d\times c}$ is a weight coefficient matrix, and $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{W}$. While the least square regression model has been successfully used in many fields, it is well known that the solution is generally overfitted to the training samples in its original form. To overcome the overfitting problem and to find a more generalized solution, a variety of its variants using different types of regularizations have been proposed [5], which can be mathematically simplified as follows:

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) + \mathcal{R}(\mathbf{W}) \tag{2}$$

where $\mathcal{R}(\mathbf{W})$ denotes a set of regularization terms.

From a machine learning point of view, a well-defined regularization term helps find a generalized solution to the objective function, and thus result in a better performance for the final goal. In this paper, we devise novel regularization terms that effectively utilize various pieces of information inherent in the observations. Note that since, in this work, we extract features from the parcellated brain areas or Regions-Of-Interest (ROIs), which are structurally or functionally related to each other, it is natural to assume that there exist relations among features. Meanwhile, if two features are highly related to each other, then it is reasonable to have the respective weight coefficients also related. However, to the best of our knowledge, none of the previous regression methods in the literature considered and guaranteed this characteristic in their solutions. To this end, we devise a new regularization term with the claim that, if some features are related to each other, the same or the similar relation is expected to be preserved between the respective weight coefficients. To utilize this '*feature-feature*' relation, we impose the relation between columns in $\mathbf{X}$ to be reflected in the relation between the corresponding rows in $\mathbf{W}$, by defining the following regularization term:

$$\mathcal{R}_1(\mathbf{W}) = \frac{1}{2}\sum_{i,j}^{d} m_{ij}\|\mathbf{w}^i - \mathbf{w}^j\|_2^2 \tag{3}$$

where m_{ij} denotes an element in the feature similarity matrix $\mathbf{M} = [m_{ij}] \in \mathbb{R}^{d\times d}$ that encodes the relation between features in the samples. Throughout this paper, we use a radial basis function kernel to measure the similarity between vectors.

Meanwhile, given a feature vector $\mathbf{x}^i$, in our joint regression and classification framework, we use a different set of weight coefficients to regress the elements in the response vector $\mathbf{y}^i$. In other words, the elements of each column in $\mathbf{W}$ are linked to the elements of each column in $\mathbf{Y}$ via feature vectors. By taking this mathematical property into account, we further impose the relation between column vectors in $\mathbf{W}$ to be similar

to the relation between the respective target response variables (i.e., respective column vectors) in $\mathbf{Y}$, which we call as '*response-response*' relation:

$$\mathcal{R}_2(\mathbf{W}) = \frac{1}{2}\sum_{i,j}^{c} g_{ij}\|\mathbf{w}_i - \mathbf{w}_j\|_2^2 \tag{4}$$

where g_{ij} denotes an element in the matrix $\mathbf{G} = [g_{ij}] \in \mathbb{R}^{c\times c}$ that represents the similarity between every pair of target response variables (i.e., every pair of column vectors in $\mathbf{Y}$). Due to the algebraic operation in the least square regression, i.e., matrix multiplication, the weight coefficients are linked to the response variables via regressors, i.e., feature vectors in our work. Therefore, it is meaningful to impose the relation between a pair of weight coefficients to be similar to the relation between the respective pair of target response variables.

We can also utilize the relational information between samples, called as '*sample-sample*' relation. That is, if samples are similar to each other, then their respective response values should be also similar to each other. In this regard, we define a regularization term as follows:

$$\mathcal{R}_3(\mathbf{W}) = \frac{1}{2}\sum_{i,j}^{n} s_{ij}\|\hat{\mathbf{y}}^i - \hat{\mathbf{y}}^j\|_2^2 = \frac{1}{2}\sum_{i,j}^{n} s_{ij}\|\mathbf{x}^i\mathbf{W} - \mathbf{x}^j\mathbf{W}\|_2^2 \tag{5}$$

where s_{ij} is an element in the matrix $\mathbf{S} = [s_{ij}] \in \mathbb{R}^{n\times n}$ that measures the similarity between every pair of samples. We should note that this kind of sample-sample relation has been successfully used in many manifold learning methods [1,6]. We argue that the simultaneous consideration of these newly devised regularization terms, *i.e.,* feature-feature relation, sample-sample relation, and response-response relation, can effectively reflect the relational information inherent in observations in finding an optimal solution.

Regarding feature selection, we believe that due to the underlying brain mechanisms that determine clinical scores or a clinical label, *i.e.,* response variables, if one feature plays a role in predicting one response variable, then it also devotes to the prediction of the other response variables. So, we further impose to use the same features across the tasks of clinical scores and clinical label prediction. Mathematically, this can be formulated by a $\ell_{2,1}$-norm on $\mathbf{W}$, *i.e.,* $\|\mathbf{W}\|_{2,1}$.

Therefore, our final objective function is formulated as follows:

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) + \alpha_1\mathcal{R}_1(\mathbf{W}) + \alpha_2\mathcal{R}_2(\mathbf{W}) + \alpha_3\mathcal{R}_3(\mathbf{W}) + \lambda\|\mathbf{W}\|_{2,1} \tag{6}$$

where α_1, α_2, α_3, and λ denote control parameters of the respective regularization terms, respectively. This objective function can be efficiently optimized using the framework in [22].

It is noteworthy that unlike the previous regularization methods such as local linear embedding [10], locality preserving projection [6], predictive space aggregated regression [2], and high-order graph matching [9] that focused on the sample similarities by imposing nearby samples to be still nearby in the transformed space, the proposed method utilizes richer information inherent in the observations. Thus, it is expected that the proposed method can find a generalized solution, which can be robust to noise or outliers.

3 Experimental Analysis

We compared the performance of the proposed method and the state-of-the-art methods on a subset of the ADNI dataset. Our dataset has 202 subjects including 51 AD, 52 NC, and 99 MCI. Moreover, 99 MCI contains 43 MCI-C and 56 MCI-NC.

3.1 Image Processing and Feature Extraction

We conducted image pre-processing for MRI and PET images by the sequential application of spatial distortion, skull-stripping, and removal of cerebellum. Then, for structural MRI images, we segmented them into three different tissues: gray matter (GM), white matter (WM), and CSF. By warping Kabani *et al.* 's atlas [8] into a subject's MRI image, we further dissected the GM tissue into 93 ROIs by HAMMER[11]. We then regarded the volume of the GM tissue of each ROI as a feature. We aligned each PET image to its corresponding MRI image, and then took the average intensity of each ROI as a feature. Thus, we extracted 93 features from MR and PET images, respectively.

3.2 Experimental Setting

We considered three binary classification problems: AD vs. NC, MCI vs. NC, and MCI-C vs. MCI-NC. For MCI vs. NC, both MCI-C and MCI-NC were labeled as MCI. For each set of experiments, we used 93 MRI features or 93 PET features as regressors, and 2 clinical scores along with 1 class label as responses in the least square regression model. We employed the metrics of Correlation Coefficient (CC) and Root Mean Squared Error (RMSE) between the target clinical scores and the predicted ones in regression, and also the metrics of classification ACCuracy (ACC), SENsitivity (SEN), SPEcificity (SPE), and Area Under the receiver operating characteristic Curve (AUC) in classification.

To validate the effectiveness of the proposed method, we considered rigorous experimental conditions: (1) In order to show the validity of the feature selection strategy, we performed the tasks of regression and classification without precedent feature selection, and considered them as a baseline method. Hereafter, we use the suffix "N" to indicate that no feature selection was involved in. For example, by MRI-N, we mean that either the classification or regression was performed using the full MRI features. (2) One of the main arguments in our work is to select features that can be jointly used for both regression and classification. To this end, we compare the multi-task based method with a single-task based method, in which the feature selection was carried out for regression and classification independently. In the following, the suffix "S" manifests a single-task based method. For example, MRI-S represents single-task based feature selection on MRI features. (3) We compare with two state-of-the-art methods: High-Order Graph Matching (HOGM) [9] and Multi-Modal Multi-Task (M3T) [15]. The former used a sample-sample relation along with an ℓ_1-norm in an optimization of single-task learning. The latter used multi-task learning with an $\ell_{2,1}$-norm only to select a common set of features for the tasks of regression and classification.

Table 1. Comparison of classification performances (%) of the competing methods

Feature	Method	AD vs. NC				MCI vs. NC				MCI-C vs. MCI-NC			
		ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
MRI	MRI-N	89.5	82.7	86.3	95.3	68.3	92.6	39.2	82.5	60.3	15.5	92.3	68.7
	MRI-S	91.2	85.9	92.5	96.7	76.7	93.3	37.6	83.7	64.5	24.9	**95.8**	70.6
	HOGM	93.4	**89.5**	92.5	97.1	77.7	**95.6**	51.4	84.4	66.8	36.7	95.0	72.2
	M3T	92.6	87.2	95.9	97.5	78.1	94.5	54.0	83.1	67.1	37.7	92.0	72.5
	Proposed	**93.7**	88.6	**97.8**	**97.6**	**79.7**	94.8	**56.9**	**84.7**	**71.8**	**48.0**	92.8	**81.4**
PET	PET-N	86.2	83.5	84.8	94.8	69.0	95.0	30.8	77.9	62.2	21.6	93.1	71.3
	PET-S	87.9	85.7	90.9	94.7	73.8	96.5	36.2	78.7	65.1	31.0	**95.5**	73.5
	HOGM	91.7	91.1	92.8	95.6	74.7	96.5	43.2	79.3	66.6	35.5	**95.5**	72.4
	M3T	90.9	90.5	93.1	96.4	77.2	94.5	44.3	80.5	67.0	39.1	93.2	73.1
	Proposed	**91.8**	**91.5**	**93.8**	**96.9**	**79.2**	**97.1**	**45.3**	**80.8**	**71.2**	**47.4**	93.0	**77.6**

3.3 Classification Results

Table 1 shows the classification performances of all the competing methods. From these results, we can draw three conclusions. First, it is important to conduct feature selection on the high-dimensional features before training a classifier since the baseline methods with no feature selection, *i.e.,* MRI-N, and PET-N, reported the worst performances. Second, it is beneficial to use joint regression and classification framework, *i.e.,* multi-task learning, for feature selection. As shown in Table 1, M3T and our method, which utilized the multi-task learning, achieved better classification performances than the single-task based method. Specifically, the proposed method showed the superiority to the single-task based method, *i.e.,* MRI-S and PET-S, improving the accuracies by 2.5% (AD vs. NC), 3.0% (MCI vs. NC), and 7.3% (MCI-C vs. MCI-NC) with MRI, and by 3.9% (AD vs. NC), 10.2% (MCI vs. NC), and 9.0% (MCI-C vs. MCI-NC) with PET, respectively. Lastly, based on the fact that the best performances over the three binary classifications were all obtained by our method, we can say that the proposed regularization terms were effective to find class-discriminative features. It is worth noting that compared to the state-of-the-art methods, the accuracy enhancements by our method were 5% (vs. HOGM) and 4.7% (vs. M3T) with MRI, and 4.6% (vs. HOGM) and 4.2% (vs. M3T) with PET for MCI-C vs. MCI-NC classification, which is the most important for early diagnosis and treatment.

3.4 Regression Results

Regarding the prediction of two clinical scores of MMSE and ADAS-Cog, we summarized the results in Table 2, we can see that, similar to the classification results, the regression performance of the methods without feature selection (MRI-N and PET-N) was worse than any of the other methods with feature selection. Moreover, our method consistently outperformed the competing methods for the cases of different pairs of clinical labels.

In the regression with MRI for AD vs. NC, our method showed the best CCs of 0.669 for ADAS-Cog and 0.679 for MMSE, and the best RMSEs of 4.43 for ADAS-Cog and

Table 2. Comparison of regression performances of the competing methods in terms of Correlation Coefficient (CC) and Root Mean Square Error (RMSE)

Feature	Method	AD vs. NC				MCI vs. NC				MCI-C vs. MCI-NC			
		ADAS-Cog		MMSE		ADAS-Cog		MMSE		ADAS-Cog		MMSE	
		CC	RMSE	CC	RMSE	CC	RMSE	CC	RMSE	CC	RMSE	CC	RMSE
MRI	MRI-N	0.587	4.96	0.520	2.02	0.329	4.48	0.309	1.90	0.420	4.10	0.441	1.51
	MRI-S	0.591	4.85	0.566	1.95	0.347	4.27	0.367	1.64	0.426	4.01	0.482	1.44
	HOGM	0.625	4.53	0.598	1.91	0.352	4.26	0.371	1.63	0.435	3.94	0.521	1.41
	M3T	0.649	4.60	0.638	1.91	0.445	4.27	0.420	1.66	0.497	4.01	0.550	1.41
	Proposed	**0.669**	**4.43**	**0.679**	**1.79**	**0.472**	**4.23**	**0.500**	**1.62**	**0.589**	**3.83**	**0.603**	**1.40**
PET	PET-N	0.597	4.86	0.514	2.04	0.333	4.34	0.331	1.70	0.382	4.08	0.452	1.50
	PET-S	0.620	4.83	0.593	2.00	0.356	4.26	0.359	1.69	0.437	4.00	0.478	1.48
	HOGM	0.600	4.69	0.515	1.99	0.360	4.21	0.368	1.67	0.430	4.03	0.523	1.41
	M3T	0.647	4.67	0.593	1.92	0.447	4.24	0.432	1.68	0.520	3.91	0.569	1.45
	Proposed	**0.671**	**4.41**	**0.620**	**1.90**	**0.513**	**4.13**	**0.485**	**1.66**	**0.526**	**3.87**	**0.570**	**1.37**

1.79 for MMSE. The next best performances in terms of CCs were obtained by M3T, *i.e.,* 0.649 for ADAS-Cog and 0.638 for MMSE, and those in terms of RMSEs were obtained by HOGM, *i.e.,* 4.53 for ADAS-Cog and 1.91 for MMSE. In the regression with MRI for MCI vs. NC, our method also achieved the best CCs of 0.472 for ADAS-Cog and 0.50 for MMSE, and the best RMSEs of 4.23 for ADAS-Cog and 1.62 for MMSE. For the case of MCI-C vs. MCI-NC with MRI, the proposed method improved the CCs by 0.092 for ADAS-Cog and 0.053 for MMSE compared to the next best CCs of 0.497 for ADAS-Cog and 0.550 for MMSE by M3T. Note that the proposed method with PET also reported the best CCs and RMSEs for both ADAS-Cog and MMSE over the three regression problems, *i.e.,* AD vs. NC, MCI vs. NC, and MCI-C vs. MCI-NC.

4 Conclusions

In this work, we proposed a novel feature selection method by devising new regularization terms that consider relational information inherent in the observations for joint regression and classification in the computer-aided AD diagnosis. From our extensive experiments on the ADNI dataset, we found that the utilization of the devised three regularization terms, *i.e.,* sample-sample relation, feature-feature relation, and response-response relation, were helpful to improve the performances in the problem of joint regression and classification, outperforming the state-of-the-art methods.

Acknowledgements. This study was supported by National Institutes of Health (EB006733, EB008374, EB009634, AG041721, AG042599, and MH100217). Xiaofeng Zhu was partly supported by the National Natural Science Foundation of China under grant 61263035.

References

1. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research 7, 2399–2434 (2006)

2. Chen, T., Kumar, R., Troianowski, G.A., Syeda-Mahmood, T.F., Beymer, D., Brannon, K.: Psar: Predictive space aggregated regression and its application in valvular heart disease classification. In: ISBI, pp. 1122–1125 (2013)
3. Cho, Y., Seong, J.K., Jeong, Y., Shin, S.Y.: Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. NeuroImage 59(3), 2217–2230 (2012)
4. Duchesne, S., Caroli, A., Geroldi, C., Collins, D.L., Frisoni, G.B.: Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. NeuroImage 47(4), 1363–1370 (2009)
5. Hastie, T., Tibshirani, R., Friedman, J., Franklin, J.: The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer 27(2), 83–85 (2005)
6. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: NIPS, pp. 1–8 (2005)
7. Jie, B., Zhang, D., Cheng, B., Shen, D.: Manifold regularized multi-task feature selection for multi-modality classification in Alzheimers disease. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part I. LNCS, vol. 8149, pp. 275–283. Springer, Heidelberg (2013)
8. Kabani, N.J.: 3D anatomical atlas of the human brain. NeuroImage 7, 0700–0717 (1998)
9. Liu, F., Suk, H.-I., Wee, C.-Y., Chen, H., Shen, D.: High-order graph matching based feature selection for Alzheimer's disease identification. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part II. LNCS, vol. 8150, pp. 311–318. Springer, Heidelberg (2013)
10. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326 (2000)
11. Shen, D., Davatzikos, C.: HAMMER: hierarchical attribute matching mechanism for elastic registration. IEEE Transactions on Medical Imaging 21(11), 1421–1439 (2002)
12. Suk, H.I., Lee, S.W.: A novel Bayesian framework for discriminative feature extraction in brain-computer interfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(2), 286–299 (2013)
13. Suk, H.-I., Shen, D.: Deep learning-based feature representation for AD/MCI classification. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part II. LNCS, vol. 8150, pp. 583–590. Springer, Heidelberg (2013)
14. Wee, C.Y., Yap, P.T., Zhang, D., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K.A., Wang, L., Shen, D.: Identification of MCI individuals using structural and functional connectivity networks. Neuroimage 59(3), 2045–2056 (2012)
15. Zhang, D., Shen, D.: Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage 59(2), 895–907 (2012)
16. Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D.: Multimodal classification of Alzheimer's disease and mild cognitive impairment. NeuroImage 55(3), 856–867 (2011)
17. Zhu, X., Huang, Z., Cheng, H., Cui, J., Shen, H.T.: Sparse hashing for fast multimedia search. ACM Transactions on Information Systems 31(2), 9 (2013)
18. Zhu, X., Huang, Z., Cui, J., Shen, T.: Video-to-shot tag propagation by graph sparse group lasso. IEEE Transactions on Multimedia 13(3), 633–646 (2013)
19. Zhu, X., Huang, Z., Shen, H.T., Cheng, J., Xu, C.: Dimensionality reduction by mixed kernel canonical correlation analysis. Pattern Recognition 45(8), 3003–3016 (2012)
20. Zhu, X., Huang, Z., Yang, Y., Shen, H.T., Xu, C., Luo, J.: Self-taught dimensionality reduction on the high-dimensional small-sized data. Pattern Recognition 46(1), 215–229 (2013)
21. Zhu, X., Suk, H.I., Shen, D.: Matrix-similarity based loss function and feature selection for Alzheimer's disease diagnosis. In: CVPR (2014)
22. Zhu, X., Wu, X., Ding, W., Zhang, S.: Feature selection by joint graph sparse coding. In: SDM, pp. 803–811 (2013)

Fisher Kernel Based Task Boundary Retrieval in Laparoscopic Database with Single Video Query

Andru Putra Twinanda, Michel De Mathelin, and Nicolas Padoy

ICube, University of Strasbourg, CNRS, France
{twinanda,demathelin,npadoy}@unistra.fr

Abstract. As minimally invasive surgery becomes increasingly popular, the volume of recorded laparoscopic videos will increase rapidly. Invaluable information for teaching, assistance during difficult cases, and quality evaluation can be accessed from these videos through a video search engine. Typically, video search engines give a list of the most relevant videos pertaining to a keyword. However, instead of a whole video, one is often only interested in a fraction of the video (e.g. intestine stitching in bypass surgeries). In addition, video search requires semantic tags, yet the large amount of data typically generated hinders the feasibility of manual annotation. To tackle these problems, we propose a coarse-to-fine video indexing approach that looks for the time boundaries of a task in a laparoscopic video based on a video snippet query. We combine our search approach with the Fisher kernel (FK) encoding and show that similarity measures on this encoding are better suited for this problem than traditional similarities, such as dynamic time warping (DTW). Despite visual challenges, such as the presence of smoke, motion blur, and lens impurity, our approach performs very well in finding 3 tasks in 49 bypass videos, 1 task in 23 hernia videos, and also 1 cross-surgery task between 49 bypass and 7 sleeve gastrectomy videos.

Keywords: surgical workflow analysis, laparoscopy, time boundaries, video indexing, sliding window, Fisher kernel.

1 Introduction

Most knowledge, in the form of texts, images, and even videos, is just one keyword and a click away thanks to search engines. Because images and videos have a rich content that still cannot be fully extracted automatically by computers, their retrieval is mostly possible because of the semantic tags provided by manual annotation. However, with vast amounts of data, browsing the videos and associating them with semantic tags manually becomes tedious. This is especially true for medical data, where the skilled annotators are moreover likely to be clinicians with little time on their hands.

In this paper, we therefore propose a method that looks for the time boundaries of a task in a video based on a video query. We use a video snippet of the task as query in order to eliminate the need for semantic tags. By providing the

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 409–416, 2014.

snippet, the method automatically provides the annotation of the task in relevant videos in the database. We call this problem video indexing and focus in this work on laparoscopic videos. Such videos contain invaluable information about the execution of surgeries in various clinical configurations. By implementing this method, fellow surgeons can observe different techniques in performing a specific surgical task. Additionally, such a method can provide boundary candidates in order to automate or simplify semantic tag annotation. Processing laparoscopic videos is, however, not a trivial problem due to atypical visual challenges, such as the presence of smoke, specular reflection, motion blur, and lens impurity.

In the domain of surgical video processing, one of the most explored topics is surgical phase recognition [1,2,3]. For instance, Lalys et al. [1] presented a framework to segment high-level surgical phases from 20 videos of cataract surgeries based on visual features; Padoy et al. [2] proposed a method that uses the signals from the surgical tool to model and recognize the surgical workflow of 16 cholecystectomies; and Blum et al. [3] combined both visual features and surgical tool signals to train a classification model for segmenting 7 cholecystectomy videos. However, all of these previous works depended on a model that was trained on videos whose phases had been previously fully annotated by a human annotator. Moreover, the model required training from complete workflows, which can be an issue when the workflow is not sequential or when the videos are incomplete. In contrast, our work aims to perform the video indexing in an unsupervised manner using a single video query.

Our work is closely related to video sequence matching, which is the process of finding similarities between two video sequences. Typically, such a method is performed to find videos that are relevant to the query video [4], but not to find the time boundaries of a particular task in the video. Other related methods, such as [5], focus on action recognition and work in a supervised environment using datasets that usually contain short actions. One of the closest works to ours is [6] where Chu et. al proposed the temporal commonality discoveries (TCD) method. It is an unsupervised approach to find the time boundaries of the commonalities between two sequences using a branch and bound (B&B) optimization with histogram encoding and l_1/χ^2 bounding distance. We observed that, despite the optimization, the B&B still carries out too many evaluations, thus taking a long time to find the task boundaries for one video. Moreover, the global optimality comes at the cost of a suitable similarity measure, namely TCD enforces the use of histogram encoding. In contrast, we are particularly interested in the Fisher kernel (FK) encoding. The FK has become popular since Perronnin et. al. [7] showed that it can be understood as an extension of the bag-of-words (BOW) approach (i.e. histogram encoding). It is a generic framework that combines the benefits of generative and discriminative approaches. As presented in a recent work [8], applying FK on frame-based features is also superior to BOW due to its ability to model the variation in time in the videos.

In this work, we propose a novel coarse-to-fine temporal search to find the time boundaries using FK-based similarity. The coarse-to-fine search speeds up time boundaries discovery compared to the traditional sliding window approaches.

While coarse-to-fine approaches are popular in image processing, they have not been investigated as much in such time-series data because the considered tasks are usually already short. Furthermore, unlike TCD, our method accepts any types of feature encoding and similarity measures. We will show in particular that FK-based similarity largely outperforms traditional similarities between time-series, such as dynamic time warping (DTW) [9]. We will also show that the combination of FK with coarse-to-fine temporal search gives higher performance compared to the globally optimal TCD. We carried out extensive experiments to retrieve the time boundaries of 3 tasks (i.e. intestine stapling, intestine stitching, and fat stitching) from 49 bypass videos, and one task (i.e. net placing) from 23 hernia videos. We also carried out a cross-surgery retrieval of the intestine stitching from the bypass videos to 7 sleeve gastrectomy videos.

In summary, the contributions of this paper are three-fold: (1) we tackle the problem of automatic video indexing which, to the best of our knowledge, has not previously been addressed in the medical community; it is also very different from the shot detection and action recognition problem from the traditional computer vision community; (2) we propose a coarse-to-fine temporal search combined with Fisher-kernel based similarity and show its suitability for laparoscopic video data; and (3) we present an extensive retrieval comparison with multiple techniques and similarity measures.

2 Methodology

2.1 Frame Rejection

This step is carried out to reject irrelevant frames (e.g. blank or static images, arbitrary views outside the patient's body) from the laparoscopic videos before the feature extraction process. Supervised methods, such as [10], have been presented to tackle this problem. However, to keep the whole process completely unsupervised, we propose a simple RGB histogram thresholding. Through observation, it is apparent that the red color channel is particularly more dominant compared to the other color channels in relevant frames (i.e. views inside the patient's body). A scalar value is computed to represent each color channel histogram in such a way that if the red scalar value is in a certain range and superior to the blue and green scalars, the frame will be accepted and then processed. It was shown in [11] that this approach significantly reduced the number of frames to be processed and improved the accuracy of surgery classification for laparoscopic videos.

2.2 Feature Representation

In the field of video processing, many visual-based features have been explored, such as color information, image gradients, optical flow and spatio-temporal interest points (STIP) [12]. However, based on preliminary experiments, the color histogram is not a discriminative feature since our frames look very similar to one

another. Also, optical flow fails due to the rapid movement of the laparoscopic camera, and the STIP was observed to be very sensitive to specular reflection. Thus, we decided to extract histogram of gradients (HOG) [13] since it acts as a global descriptor for the video frames.

We encode the features using two approaches: bag-of-words (BOW) and fisher kernel (FK). We use the typical BOW approach using k-means clustering to build the vocabulary with a hard data-to-cluster assignment. For the fisher kernel, we represent the vocabulary as a Gaussian Mixture Model (GMM) with K Gaussians. As repeatedly suggested in many works such as [8], we also tried to reduce the dimensionality of HOG using principal component analysis (PCA) for the FK representation. However, we found during preliminary experiments that the dimensionality reduction did not bring any improvements to the overall precision and recall, so we keep the original dimensions of the features.

2.3 Video Sequence Similarity

Given two video sequences $\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \dots \mathbf{a}_{nA} \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} \mathbf{b}_1 \dots \mathbf{b}_{nB} \end{bmatrix}$ where nA is the number of frames in video $\mathbf{A}$ (respectively nB and $\mathbf{B}$) and $\mathbf{a}_i$ is the vector representation of the i-th frame in video $\mathbf{A}$ (respectively $\mathbf{b}_i$ and $\mathbf{B}$), the similarity between the two is defined as either $S(\mathbf{A}, \mathbf{B}) = DTW(\mathbf{A}, \mathbf{B})$ or $S(\mathbf{A}, \mathbf{B}) = D_d^e(\mathbf{A}, \mathbf{B})$. $DTW(\mathbf{A}, \mathbf{B})$ computes the similarity between $\mathbf{A}$ and $\mathbf{B}$ using DTW, while $D_d^e(\mathbf{A}, \mathbf{B})$ computes the similarity using encoding $e \in \{\text{BOW, FK}\}$ and distance d. In this paper, we consider vector distances (i.e. l_1 and l_2), histogram distance (i.e. χ^2), and also mutual information (MI).

2.4 Boundary Search

We represent our query and target videos respectively as matrices $\mathbf{Q} = \begin{bmatrix} \mathbf{q}_1 \dots \mathbf{q}_m \end{bmatrix}$ and $\mathbf{R} = \begin{bmatrix} \mathbf{r}_1 \dots \mathbf{r}_n \end{bmatrix}$ where $m < n$. The problem of video indexing is to find the best time interval $[b, e] \subseteq [1, n]$ in the target video, such that

$$(b^*, e^*) = \arg\min_{b,e} S(\mathbf{Q}[1, m], \mathbf{R}[b, e]), \tag{1}$$

where $\mathbf{R}[b, e] = \begin{bmatrix} \mathbf{r}_b \dots \mathbf{r}_e \end{bmatrix}$ denotes the subsequence of $\mathbf{R}$ that begins from frame b and ends in frame e, hence $\mathbf{Q}[1, m] = \mathbf{Q}$.

We initialize our algorithm by temporally partitioning the target video $\mathbf{R}$ into L overlapping segments with the size of m. The amount of overlapping depends on the predefined time step $s = \alpha m$, where $0 < \alpha < 1$ to ensure overlapping. For a target video $\mathbf{R}$, this partitioning then defines $\mathbf{R}_i = \mathbf{R}[b_i, e_i]$ where $i \in \{1, L\}$, $b_1 = 1$, $e_i = b_i + m - 1$, and $b_j = b_{j-1} + s, j \in \{2, L\}$.

We find the most similar segment to the query by computing the similarity:

$$i^* = \arg\min_i S(\mathbf{Q}, \mathbf{R}_i) \tag{2}$$

Taking $\mathbf{R}_{i^*}$ as our initial segment, we find the time interval by refining the boundary through boundary shrinking and expansion. This is carried out since

the size of the queried task in video $\mathbf{R}$ is not necessarily equal to the query size m. By considering a step σ, we compute the distance of $\mathbf{Q}$ to four possible segments: $\mathbf{R}[b_{i^*} + \sigma, e_{i^*}]$, $\mathbf{R}[b_{i^*} - \sigma, e_{i^*}]$, $\mathbf{R}[b_{i^*}, e_{i^*} + \sigma]$, and $\mathbf{R}[b_{i^*}, e_{i^*} - \sigma]$. Next, we choose the best segment among the four possible ones. This process is repeated until the similarities between the query and all four possible segments are less than between the query and the initial segment.

2.5 Coarse-to-Fine Approach

In order to improve the computational time, we propose to use a coarse-to-fine approach. Both query and target videos are downsampled N times by a factor of 2. The search algorithm begins at the lowest resolution. Next, we limit our search on the higher resolutions based on the result (i.e. the time boundaries) from the search at lower resolution. This process is repeatedly done until the highest resolution is reached. This way, the total number of comparisons is reduced since we get the rough estimation of the time boundaries from the lower resolutions.

3 Experimental Results

We conducted experiments to retrieve 4 tasks: intestine stitching (IStit), intestine stapling (IStap), fat stitching (FStit) and net placing (NP). Our dataset consists of 79 surgeries performed by 8 surgeons. The details of the tasks and surgeries are shown in Table 1-b. There are only 45 videos of 49 videos for the task IStap due to incomplete recordings. For quantitative analysis, we manually annotated the time boundaries of the tasks in all videos. To evaluate the method, we performed random testing for bypass and hernia surgeries by searching 4 random queries within the remaining videos. We repeated this process 5 times. The underlying assumption was that the task was present in every target video. Methods, such as in [4], that retrieve relevant videos can be used to determine whether the task is present in the target videos. This is however not the focus of this work.

In Table 1-c, the full experimental setup is shown and the explanation of the naming conventions is given in Table 1-a. We trained BOW and GMM dictionaries with respectively 500 and 50 words. We set $\alpha = 0.1$ to get high overlapping segments and a large refinement step $\sigma = 5$. For TCD configuration, we used the χ^2 bounding function since it was shown in [6] to provide the best result. This was also confirmed in our preliminary experiments. We downsampled the videos to 0.5 frame per second (fps) for the best performance-time trade-off. For our method, we carried out extensive experiments to observe the effects of various parameters. Due to limited space, we only show the most significant results.

For evaluation, we use the method proposed in [6]. Given the ground truth of the time boundaries $\hat{T} = [\hat{b}, \hat{e}]$, the evaluation of the estimated time boundaries $T = [b, e]$ is done by computing precision $\phi_p = \frac{T \cap \hat{T}}{T}$ and recall $\phi_r = \frac{T \cap \hat{T}}{\hat{T}}$. We consider a boundary estimation to be correct in terms of precision if $\phi_p > 0.5$, meaning more than half of the frames in the estimated time boundaries is in the ground true boundaries. We also apply the similar thresholding for recall.

Table 1. Experimental setup: (a) the naming conventions for the experimental configurations; (b) the details of the tasks and surgeries, including number of surgeries and the mean ± std of the task and surgery length; and (c) the configuration of experiments complying to the conventions defined in (a), except for TCD.

Code	Description
C	Coarse-to-fine approach
B	BOW representation
F	FK representation
L	Low resolution
H	High resolution
R	Refinement

(a)

Task	Surgery	#Surg.	Avg. Len. (min.)	
			Task	Surg
IStit	Bypass	49	14 ± 5	111 ± 27
	Sleeve Gast.	7	25 ± 8	109 ± 16
FStit	Bypass	49	6 ± 2	111 ± 27
IStap	Bypass	45	7 ± 5	114 ± 26
NP	Hernia	23	4 ± 2	50 ± 25

(b)

ID	FPS	N
TCD	0.5	-
CBH-DTW	2.5	3
CBH-MI	2.5	3
CBH-χ^2	2.5	3
FL-l_1	0.5	1
FL-l_2	0.5	1
FH-l_1	2.5	1
FH-l_2	2.5	1
CFH-l_1	2.5	3
CFH-l_2	2.5	3
CFHR-l_1	2.5	3
CFHR-l_2	2.5	3

(c)

Comparison with TCD. From multiple search configurations shown in Table 1-c, we show in Table 2-a the best configurations (with the highest $F = 2 \cdot \frac{\text{rec} \cdot \text{prec}}{\text{rec} + \text{prec}}$) of our video indexing method compared to TCD. Compared to TCD, our method is significantly faster since it does less evaluations. Note that the best results are all obtained from the coarse-to-fine configurations. Thus, not only does the coarse-to-fine approach decrease the number of evaluations, but it also improves the performance of the method. This is possible because less noise is present in lower resolution, giving better initialization at the higher resolution.

For all events, our method gives the same or higher precision compared to TCD. The recall is only slightly decreased for IStit task and largely increased for all the other tasks. Out of 5 task retrievals, 4 are obtained using coarse-to-fine approaches with FK, 3 of which use the boundary refinement method.

It can be seen that the method tends to fail at retrieving IStap and NP tasks. Compared to stitching tasks, these tasks have higher variability in terms of the sub-task execution. However, the precision and recall are still much higher than TCD and chance. In summary, the complete approaches that we propose, namely CFHR-l_1 and CFHR-l_2, have higher average F-measure (respectively 35.29 and 35.36) over all retrievals compared to TCD (34.88).

Effect of Resolution, Feature Representation and Choice of Similarity Measures. To observe the effect of various parameters, we show the complete results from IStit task retrieval in Bypass videos using queries from Bypass videos (B→B) in Table 2-b. Since our method performs really fast, we have the possibility to increase the data resolution up to full resolution. In preliminary

Table 2. Experimental results. (a) Comparison of precision, recall, and mean ± std of execution time between TCD and our boundary search method. Q and T respectively stands for query and target. We use the first letter of the surgery to ID them, thus B, S, and H respectively represents bypass, sleeve gastrectomy, and hernia. For instance, B→S means looking for a task based on video query from bypass surgery in sleeve gastrectomy videos. (b) Precision and recall comparison of our method for IStit B→B.

Task	Q→T	TCD			Best of ours			
		Prec (%)	Rec (%)	Time (s)	ID	Prec (%)	Rec (%)	Time (s)
IStit	B→B	70.91	**78.26**	33 ± 21	CFHR-l_1	**78.57**	77.85	4.4 ± 3.3
	B→S	58.57	**22.14**	45 ± 24	CFHR-l_1	**59.28**	19.28	1.9 ± 0.5
FStit	B→B	37.44	38.19	21 ± 12	CBH-MI	**53.51**	**45.61**	0.3 ± 0.2
IStap	B→B	10.02	12.44	26 ± 17	CFH-l_2	**23.44**	**26**	2.4 ± 1.2
NP	H→H	19.47	18.52	9 ± 8.2	CFHR-l_2	**38.04**	**31.95**	2.2 ± 1.2

(a)

Task	Eval.	CBH			FL		FH		CFH		CFHR	
		DTW	MI	χ^2	l_1	l_2	l_1	l_2	l_1	l_2	l_1	l_2
IStit, B→B	Prec	40.91	70.12	56.42	77.65	58.36	76.12	58.16	77.95	58.57	78.57	58.77
	Recall	41.83	71.3	55.51	78.06	57.85	77.85	58.26	78.36	57.95	77.85	56.53

(b)

results, we observed that 2.5 fps gave the best performance-cost trade-off. Having data with higher resolution means that the we can retrieve the boundary more precisely. However, searching directly on the high resolution may cause wrong initialization due to the presence of noise. Thus, most of the time, the improvement from using higher resolution is obtained after the coarse-to-fine approach. This can be observed from the FL, FH and CFH configurations.

As expected, the FK representation performs better in most cases than the BOW approach, which confirms the conclusions of previous works [14,8]. In terms of similarity measures, our FK-based approach significantly outperforms the usual similarity on sequence, DTW, by over 35%. However, it can be noticed that the FK is sensitive to the distance function (i.e. l_1- and l_2-distance).

4 Conclusions

In this paper, we present a video indexing method for a laparoscopic video database using Fisher kernel based similarities. This method performs significantly better than DTW, the standard similarity measure between sequences. In addition, our method is more flexible than TCD as it can be adapted to any feature representation and similarity measure. We also compared our method with the globally optimal TCD and showed large improvements in most cases by using more suited similarities. Furthermore, we demonstrated that in addition to decreasing the number of evaluations, the coarse-to-fine approach does not impede the method's performance. In future work, we plan to try our approach on bigger datasets containing a greater number of tasks and explore more possibilities for cross-surgery retrievals. We also plan to further investigate the use

of other distance functions for FK in order to find a function that will work uniformly well in all cases.

Acknowledgements. This work was supported by French state funds managed by the ANR within the Investissements d'Avenir program under references ANR-11-LABX-0004 (Labex CAMI), ANR-10-IDEX-0002-02 (IdEx Unistra) and ANR-10-IAHU-02 (IHU Strasbourg). The authors would like to thank the IRCAD audio-visual team for their help in generating the dataset.

References

1. Lalys, F., Riffaud, L., Bouget, D., Jannin, P.: A framework for the recognition of high-level surgical tasks from video images for cataract surgeries. IEEE Trans. Biomed. Engineering 59(4), 966–976 (2012)
2. Padoy, N., Blum, T., Ahmadi, S.A., Feussner, H., Berger, M.O., Navab, N.: Statistical modeling and recognition of surgical workflow. Medical Image Analysis 16(3), 632–641 (2012)
3. Blum, T., Feußner, H., Navab, N.: Modeling and segmentation of surgical workflow from laparoscopic video. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part III. LNCS, vol. 6363, pp. 400–407. Springer, Heidelberg (2010)
4. Chen, L.H., Chin, K.H., Liao, H.Y.: An integrated approach to video retrieval. In: 19th Australasian Database Conference. CRPIT, vol. 75, pp. 49–55. ACS (2008)
5. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: ICCV (2013)
6. Chu, W.-S., Zhou, F., De la Torre, F.: Unsupervised temporal commonality discovery. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 373–387. Springer, Heidelberg (2012)
7. Perronnin, F., Dance, C.R.: Fisher kernels on visual vocabularies for image categorization. In: CVPR (2007)
8. Mironica, I., Uijlings, J., Rostamzadeh, N., Ionescu, B., Sebe, N.: Time matters! capturing variation in time in video using fisher kernels. ACM Multimedia (2013)
9. Sakoe, H.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. on Acoustics, Speech, and Signal Processing 26, 43–49 (1978)
10. Atasoy, S., Mateus, D., Meining, A., Yang, G.Z., Navab, N.: Endoscopic video manifolds for targeted optical biopsy. IEEE Trans. Med. Imaging 31(3), 637–653 (2012)
11. Twinanda, A.P., Marescaux, J., De Mathelin, M., Padoy, N.: Towards better laparoscopic video database organization by automatic surgery classification. In: Stoyanov, D., Collins, D.L., Sakuma, I., Abolmaesumi, P., Jannin, P. (eds.) IPCAI 2014. LNCS, vol. 8498, pp. 186–195. Springer, Heidelberg (2014)
12. Laptev, I.: On space-time interest points. Int. J. Comput. Vision 64(2-3), 107–123 (2005)
13. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. 886–893 (2005)
14. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: BMVA, pp. 76.1–76.12 (2011)

Multi-scale Analysis of Imaging Features and Its Use in the Study of COPD Exacerbation Susceptible Phenotypes

Felix J.S. Bragman[1], Jamie R. McClelland[1], Marc Modat[1], Sébastien Ourselin[1], John R. Hurst[2], and David J. Hawkes[1]

[1] Centre for Medical Image Computing, University College London, UK
[2] Centre for Inflammation and Tissue Repair, University College London, UK

Abstract. We propose a novel framework for exploring patterns of respiratory pathophysiology from paired breath-hold CT scans. This is designed to enable analysis of large datasets with the view of determining relationships between functional measures, disease state and the likelihood of disease progression. The framework is based on the local distribution of image features at various anatomical scales. Principal Component Analysis is used to visualise and quantify the multi-scale anatomical variation of features, whilst the distribution subspace can be exploited within a classification setting. This framework enables hypothesis testing related to the different phenotypes implicated in Chronic Obstructive Pulmonary Disease (COPD). We illustrate the potential of our method on initial results from a subset of patients from the COPDGene study, who are exacerbation susceptible and non-susceptible.

1 Introduction

Exacerbations of Chronic Obstructive Pulmonary Disease (COPD) are defined as a sudden worsening of symptoms, which accelerate the decline in lung function leading to an increased risk of mortality. Understanding their pathophysiology is critical for predicting the patients at greatest risk of hospitalisation. Recent work suggests that the frequency of exacerbations is a distinct phenotype [1]. This is described as an exacerbation susceptible phenotype, where a patient may exhibit distinct physiological patterns resulting in an intrinsic susceptibility.

Recent studies have suggested a potential link between changes in lung structure, function and exacerbations. A correlation between the progression of emphysema and the presence of exacerbations has been observed [2] whilst pulmonary arterial enlargement has been seen to be a related factor [3]. Further, regional ventilation defects have been observed prior to acute exacerbations [4]. These suggest a dependence between abnormalities in lung structure, the distribution of disease and exacerbations, which motivates our algorithm.

There is a growing interest in employing machine learning for the study and diagnosis of COPD. Classifiers are frequently trained with scalar values representing the whole lung [5] or individual lobes [6]. This ignores the spatial distribution of disease; which may be a signature of various COPD phenotypes.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 417–424, 2014.

We propose a novel framework for the analysis of lung pathophysiology. We hypothesise that the spatial distribution of disease is a discriminating factor in the presence of pathology. Our method is based on the measurement of image features representing the biomechanics and density of tissue, using a sliding box window at various anatomical scales. This is to deal with the bifurcating nature of the respiratory system. We apply it to the study of exacerbation susceptible and non-susceptible patients. The distributions measured at multiple scales are exploited to investigate differences between subtypes whilst classifying for the first time, those at greatest risk of further exacerbations.

2 Method

2.1 Non-rigid Registration

The NiftyReg registration platform[1] [7] is employed to find the spatial mapping between the lung at full inhalation (Ω^*) and end exhalation (Ω). This is performed using a stationary velocity field, parameterised through a cubic B-spline interpolation. The Local Normalised Cross Correlation (LNCC) drives the registration whilst the bending energy of the velocity field is used as the regularisation. The registration is performed by considering only the lungs, delineated by segmented masks. The background volume is set to 0 Hounsfield Units upon which the masks are diluted to include a 0 HU border within the lung volume.

2.2 Feature Extraction

The transformation $\varphi : \Omega \rightarrow \Omega^*$, resulting from the registration serves to map each coordinate $x \in \Omega$ to $x^* \in \Omega^*$, such that the position of voxels at expiration ($x \in \Omega$) is known within the inspiratory phase ($x^* \in \Omega^*$). Biomechanical and density-based feature sets are derived using the information embed within φ.

Biomechanical Feature Set. To quantify the transformation φ, we consider the deformation gradient tensor $\mathbf{F}$, which is defined as $\nabla_{x^*}\varphi(x)$. We derive 3 features from $\mathbf{F}$ to capture the respiratory process; the Jacobian determinant ($\det(\mathbf{F})$) and the first 2 moments of the distribution of the eigenvalues of the Lagrangian strain tensor ($\mathbf{E}$). The Jacobian determinant is defined as

$$\det(\mathbf{F}) = \det(\nabla_{x^*}\varphi(x)) \tag{1}$$

and measures the fractional volume change of voxels. The Lagrangian Strain Tensor $\mathbf{E}$ is derived from $\mathbf{F}$, by considering the Right Cauchy-Green Strain ($\mathbf{C}$)

$$\mathbf{C} = \mathbf{F}^\top\mathbf{F}, \;\; \mathbf{F} = \mathbf{R}\,\mathbf{U}, \;\; \mathbf{F}^\top\mathbf{F} = \mathbf{U}^\top\,\mathbf{R}^\top\,\mathbf{R}\,\mathbf{U} = \mathbf{U}^\top\mathbf{U} \;\; .$$

We are interested in analysing the stretches captured by $\mathbf{F}$. The tensor $\mathbf{C}$ results from a polar decomposition of $\mathbf{F}$, where the rotation component $\mathbf{R}$ is discarded

[1] `http://sourceforge.net/projects/niftyreg`

by considering its orthogonal properties. The tensor $\mathbf{C}$ is thus rotation free, solely containing information about the stretches $\mathbf{U}$. The computation of the Lagrangian Strain Tensor ($\mathbf{E}$) follows

$$\mathbf{E} = \frac{1}{2}\left(\mathbf{C} - \mathbf{I}\right) \quad . \tag{2}$$

We derive the principal strains ($\lambda = \{\lambda_i \,|\, i = 1, 2, 3\}$) via an eigen-decomposition of $\mathbf{E}$. The trace ($\sum \lambda$), provides an overall measure of the magnitude of tissue strain whilst the variance ($\mathrm{Var}(\lambda)$) characterises anisotropy in the strain profile.

Density-Based Feature Set. The transformation φ allows us to compute corresponding measures of voxel density (HU) at inspiration ($\mathrm{I_{ins}}$) and expiration ($\mathrm{I_{exp}}$). We consider the distribution of HU in $\mathrm{I_{ins}}$ and $\mathrm{I_{exp}}$ and 2 scalar values; the percentage of emphysema ($\%\mathrm{LAA_{ins}} - 950\mathrm{HU}$) and gas trapping ($\%\mathrm{LAA_{exp}} - 856\mathrm{HU}$). The $\%\mathrm{LAA_{ins/exp}}$ metrics are computed as follows:

$$\%\mathrm{LAA_{ins}} - 950\mathrm{HU} = \frac{\sum_{x^* \in \Omega^*} \left(\mathrm{I_{ins}}(x^*) < -950\right)}{\sum_{x^* \in \Omega^*} x^*} \tag{3a}$$

and

$$\%\mathrm{LAA_{exp}} - 856\mathrm{HU} = \frac{\sum_{x^* \in \Omega'} \left((\mathrm{I_{exp}}(x) \circ \varphi) < -856\right)}{\sum_{x^* \in \Omega'} x^*} \quad . \tag{3b}$$

They are expressed as the percentage of voxels below -950 HU and -856 HU within $\mathrm{I_{ins}}$ and $\mathrm{I_{exp}}$. Within our framework, all features are calculated within local neighbourhoods across the lung, which is discussed below in Sect. 2.3.

2.3 Multi-scale Analysis of Imaging Features

Feature Distributions. We propose to sample the local variation of features (f_k) to quantify their distribution across the lung. This is performed by considering histograms ($h_i(f_k; x_j, \phi_i)$) of the local distributions of f_k. Each local feature distribution is centered at a voxel x_j ($j = 1 \cdots J$) within a neighbourhood ω governed by the scale ϕ_i, where $i = 1 \cdots n$ and j is the j^{th} sampled neighbourhood. Thus, distributions at increasing scales of analysis (ϕ_i) can be computed (Fig. 1). The histograms are modelled by the first 4 statistical moments and the median. The feature f_k within ω centered at x_j is defined by:

$$\begin{aligned} H^j\left(f_k(x_j)\right) = \{&\mu(h_1)\ \nu(h_1)\ \sigma(h_1)\ \gamma_1(h_1)\ \gamma_2(h_1)\ \cdots \\ &\mu(h_n)\ \nu(h_n)\ \sigma(h_n)\ \gamma_1(h_n)\ \gamma_2(h_n)\} \end{aligned} \tag{4}$$

where μ is the mean, ν the median, σ the variance, γ_1 the skewness and γ_2 is the kurtosis. A patient-specific matrix ($\mathbf{H}_p$, $p = 1 \cdots P$) is created such that

$$\mathbf{H}_p = \begin{bmatrix} H^1\left(f_1(x_j)\right) & \cdots & H^1\left(f_k(x_j)\right) & \%\mathrm{LAA^1_{ins/exp}}(x_j)\,\forall\phi \\ \vdots & & \vdots & \vdots \\ H^J\left(f_1(x_J)\right) & \cdots & H^J\left(f_k(x_J)\right) & \%\mathrm{LAA^J_{ins/exp}}(x_J)\,\forall\phi \end{bmatrix} \quad . \tag{5}$$

The Jacobian determinant ($\det(\mathbf{F})$), the trace ($\sum \lambda$) and variance ($\mathrm{Var}(\lambda)$) of the strain eigenvalues and the voxel densities in $\mathrm{I}_{\mathrm{ins}}$ and $\mathrm{I}_{\mathrm{exp}}$ are modelled locally across the lung ($k = 5$). We incorporate the $\%\mathrm{LAA}-950\mathrm{HU}$ and $\%\mathrm{LAA}-856\mathrm{HU}$ for all ϕ_i, leading to $27n$ features per x_j. The number of sampled regions is determined by the sampling frequency of x_j at the finest scale (ϕ_1).

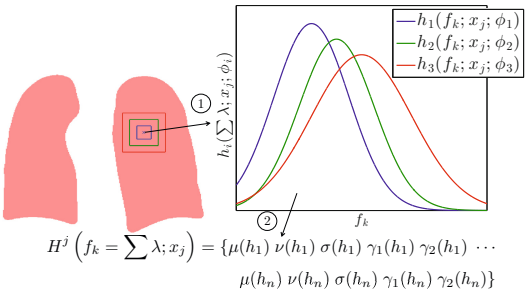


Fig. 1. Illustration of the framework. 1) A feature f_k (e.g. $\sum \lambda$) at x_j is sampled at $n = 3$ scales, leading to 3 local histograms $h_i(f_k; x_j, \phi_i)$. 2) Statistical moments and the median of $h_i(f_k; x_j, \phi_i)$ are calculated for all ϕ_i, leading to the set $H^j\left(f_k(x_j)\right)$.

Statistical Analysis of Features

Hypothesis testing using $\boldsymbol{H_p}$. Analysis of the distribution of values contained within each $\mathbf{H}_p$ allows hypotheses of changes in the global nature of local features to be made. For instance, consider the distribution of the variance of $\det(\mathbf{F})$ at all x_j. Each value demonstrates the local variation in volume change. The distribution of this measure across the lung will illustrate how the local variation is expressed, which may vary across subtypes. This facilitates a direct comparison of patient-specific distributions across phenotypes.

Principal Component Analysis of $\boldsymbol{X}$. We are interested in modelling the distribution of parameters across the studied population. We apply PCA on $\mathbf{X} = \left[\mathbf{H}_1^\top \cdots \mathbf{H}_P^\top\right]$. This seeks a low-dimensional projection ($d << 27n$) of $\mathbf{X}$, where the variance of the projected features is maximised. The entries of $\mathbf{X}$ are representative of the local histogram features measured at multiple scales. PCA of $\mathbf{X}$ allows one to compute the component scores within each neighbourhood defined by x_j. Thus, the computed scores can be projected to the image space to assess their distribution across the lung. Since the component scores are linear projections of the features measured at various scales, they will capture potential fractal properties in line with the nature of the lung anatomy. The distribution of the principal component scores can be analysed to model patient-specific distributions by computing their respective mean and variance. Thus, phenotype-specific distributions can be estimated to produce a clinically meaningful classifier. Importantly, classification in the PCA subspace prevents overfitting as PCA removes colinearity in the features.

3 Experiments and Results

3.1 Clinical Data

Inhale and exhale breath-hold CT images from the COPDGene study [8] were used. CT scans were acquired from multi-detector CT scanners, at full inspiration (200mAs) and at the end of normal expiration (50 mAs) with resolutions approximately equal to 0.66mm x 0.66mm x 0.73mm [8].

We tested our framework on $P = 20$ subjects with a GOLD 3 severity stage exhibiting $f = 0$ ($n = 10$) or $f \geq 6$ ($n = 10$) exacerbations per year. GOLD 3 patients were chosen due to their low variation in FEV_1. We chose two extreme sets ($f = 0$ and ≥ 6) to gauge the applicability of our framework in discriminating these phenotypes. The patient sets had a mean age of 60.2 and 67.5, a mean FEV_1%predicted of 42.1 and 40.5 and a mean $^{\text{FEV}_1}/_{\text{FVC}}$ ratio of 42.4 and 47.2.

3.2 Algorithm Parameters

Prior to the registration, the masks were dilated with a sphere of 3 voxel radius. An analysis of the registration parameters was performed; demonstrating robustness in the registration to small parameter changes. The standard deviation of the LNCC Gaussian kernel was set to 3^3 voxels, whilst the weighting of the regularisation was 0.05% of the overall optimised cost function. The finest control point spacing of the B-spline grid was set to 5 voxels along each axis. After registration, the inhale lung mask was eroded by a spherical element with a 7 voxel radius. This was performed to ignore regions prone to discontinuities and which experience an extreme degree of motion. We performed the sampling using a cubic box window at scales 10, 20 and 30 mm^3 ($n = 3$), which is consistent with the size of the secondary pulmonary lobule. A sampling frequency of 10mm was used yielding approximately $7,500$ regions per lung. We ignored regions at all scales where 50% of the voxels fell outside the lung mask.

3.3 Multi-scale Analysis of Imaging Features

We investigated feature distributions at the 3 scales using $\mathbf{H}_p$. We calculated the mean and standard deviation of each feature within $\mathbf{H}_p$ for all 3 scales. This provided two patient-specific distributions of values for each feature. We performed a two-sample t-test for each subtype mean and standard deviation set to determine discriminating factors between both subtypes.

A significant difference in the mean of $\sigma\left(\det(\mathbf{F})\right)$ ($.12 \pm .01$ and $.21 \pm .02$) at all scales of analysis was found ($p < .03$). The feature $\sigma\left(\det(\mathbf{F})\right)$ illustrates the variation in local volume change. The lower variation seen by the exacerbation susceptible group suggested that they exhibit a more homogeneous pattern in their volume change. No significance was seen in the standard deviation of $\sigma\left(\det(\mathbf{F})\right)$ ($p < .20$). We observed a marked difference ($p < .05$) in the mean ($.15 \pm .02$ and $.27 \pm .04$) and standard deviation ($.12 \pm .01$ and $.22 \pm .02$) of $\sigma\left(\sum(\lambda)\right)$ at all scales. This insinuated that for the susceptible group, the anisotropy in the

magnitude of local tissue strain and its variation throughout the lung is more homogeneous compared to the non-susceptible patients. These suggested a possible distinction in physiological patterns, which were exploited in the classification.

Results from the PCA of matrix $\mathbf{X}$ corroborated the above, displaying evidence of distinct feature distributions across subtypes. (Figs. 2 and 3). Figure 2 illustrates 2 patient-specific principal component distributions for each subtype. These are characteristic of the phenotype distributions and are mostly consistent across each group. As the component scores are a linear projection of the features, Figure 2 suggests that there is a consistent physiological pattern per subtype. This is illustrated by a variation in the heterogeneity of the scores as observed in the analysis of $\mathbf{H}_p$. This reinforces the notion of phenotype-specific distributions and the discriminating power of the distribution of disease.

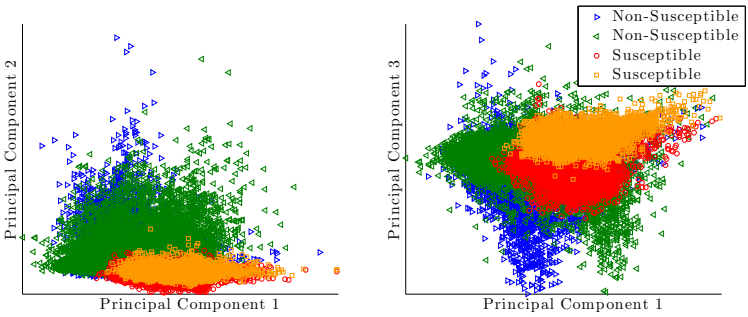
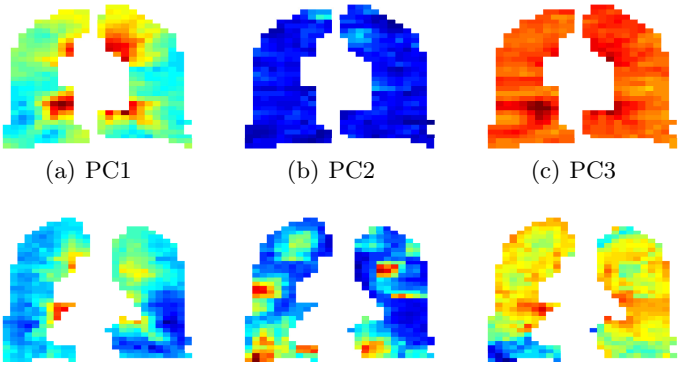


Fig. 2. An example of the multi-scale principal component distributions for each sampled neighbourhood x_j for 4 patients (2 susceptible and 2 non-susceptible patients). The first 3 principal components explain $\approx 55\%$ of the variance of $\mathbf{X}$.

As the principal component scores were computed per sampled region (x_j), we were able to couple them with their respective anatomical location (Fig. 3). This displays varying patterns in the physiology of the lung, consistently within and across subtypes. As the principal components aimed to fully explain the lung macrostructure and the deformation captured within $\mathbf{H}_p$, these maps display a novel way of viewing how lung physiology differs with the COPD phenotype and the frequency of exacerbations.

3.4 Classification of COPD exacerbation-susceptible patients

We aimed to classify exacerbation susceptible and non-susceptible patients based on the hypothesis that global and local patterns of disease differed across subtypes. This was shown in Fig. 2 and 3, where a rise in feature homogeneity coincided with exacerbation susceptibility. We performed the classification on the feature projections using the mean and the variance of the principal component scores as features. The set explaining 90% (17/81) of the variance of $\mathbf{X}$ was chosen.



(a) PC1 (b) PC2 (c) PC3

Fig. 3. Projection of the first 3 principal components of the multi-scale PCA into the image space. Coronal slice is at the mid-section. Top row: exacerbation susceptible phenotype. Bottom row: exacerbation non-susceptible phenotype.

A leave-one-out cross validation (LOOCV) was employed to test the classifier. LOOCV iteratively selects one patient ($\mathbf{H}_{unseen}$) as the testing data whilst the remaining are used for training. We assumed independence amongst each training set during the LOOCV to calculate accuracy and precision rates. For the classification, we projected $\mathbf{H}_{unseen}$ into the principal component space of $\mathbf{X}_{P-1}$ and used the mean and variance of the principal component scores as features. We used Support Vector Machines (SVM) as a classifier with a Gaussian radial basis function kernel $\sigma = 2.25$ and a soft-margin constant $C = 0.5$. Our framework has the unique ability to classify an unseen patient as either exacerbation susceptible ($f \geq 6$) or non-susceptible ($f = 0$) with a total accuracy of 75% (Table 1). This supports the applicability of our framework towards determining relationships between the distribution of disease with the clinical outcome.

Table 1. Classification results using Leave One-Out Cross Validation

	Susceptible	Non-Susceptible	Total
Classification accuracy (%)	80	70	75 ± 7.5

4 Conclusions

We have presented a novel framework for investigating global and local patterns of lung pathophysiology. The applicability of our framework in determining relationships between functional measures and the severity of disease has been shown, through an analysis of the exacerbation susceptible phenotype. Analysis of the local feature distributions displayed significant differences in the nature of lung function across subtypes. This translated to subtype-specific distributions

after dimensionality reduction, suggesting an intrinsic physiological behaviour attributed to both sets of patients. The main limitation of our work is due to the lack of patients analysed. We aim to include a larger population of patients to better demonstrate the clinical applicability of our work. This will allow us to correctly evaluate the performance of our classifier, and the consistency and utility of the derived feature distributions. Moreover, we intend to construct anatomical atlases to perform regional inter-patient statistics to investigate whether the spatial location of disease provides a further dimension to the analysis.

Acknowledgements. This work was supported by the EPSRC (EP/H046410/1 and EP/K502959/1) and the National Institute for Health Research University College London Hospitals Biomedical Research Centre. This research used data generated by the COPDGene study (phs000179.v3.p2), which was supported by NIH grants U01HL089856 and U01HL089897.

References

1. Hurst, J.R., Vestbo, J., Anzueto, A., Locantore, N., Müllerova, H., Tal-Singer, R., Miller, B., Lomas, D.A., Agusti, A., Macnee, W., Calverley, P., Rennard, S., Wouters, E.F.M., Wedzicha, J.A.: Susceptibility to exacerbation in chronic obstructive pulmonary disease. The New England Journal of Medicine 363(12), 1128–1138 (2010)
2. Tanabe, N., Muro, S., Hirai, T., Oguma, T., Terada, K., Marumo, S., Kinose, D., Ogawa, E., Hoshino, Y., Mishima, M.: Impact of exacerbations on emphysema progression in chronic obstructive pulmonary disease. American Journal of Respiratory and Critical Care Medicine 183(12), 1653–1659 (2011)
3. Wells, J.M., Washko, G.R., Han, M.K., Abbas, N., Nath, H., Mamary, A.J., Regan, E., Bailey, W.C., Martinez, F.J., Westfall, E., Beaty, T.H., Curran-Everett, D., Curtis, J.L., Hokanson, J.E., Lynch, D.A., Make, B.J., Crapo, J.D., Silverman, E.K., Bowler, R.P., Dransfield, M.T.: Pulmonary arterial enlargement and acute exacerbations of COPD. The New England Journal of Medicine 367(10), 913–921 (2012)
4. Kirby, M., Kanhere, N., Etemad-Rezai, R., McCormack, D.G., Parraga, G.: Hyperpolarized Helium-3 magnetic resonance imaging of chronic obstructive pulmonary disease exacerbation. Journal of Magnetic Resonance Imaging 37(5), 1223–1227 (2013)
5. Bodduluri, S., Newell, J.D., Hoffman, E.A., Reinhardt, J.M.: Registration-based lung mechanical analysis of chronic obstructive pulmonary disease (COPD) using a supervised machine learning framework. Academic Radiology 20(5), 527–536 (2013)
6. Murphy, K., Pluim, J.P.W., van Rikxoort, E.M., de Jong, P.A., de Hoop, B., Gietema, H.A., Mets, O., de Bruijne, M., Lo, P., Prokop, M., van Ginneken, B.: Toward automatic regional analysis of pulmonary function using inspiration and expiration thoracic CT. Medical Physics 39(3), 1650–1662 (2012)
7. Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S.: Fast free-form deformation using graphics processing units. Computer Methods and Programs in Biomedicine 98(3), 278–284 (2010)
8. Regan, E.A., Hokanson, J.E., Murphy, J.R., Make, B., Lynch, D.A., Beaty, T.H., Curran-Everett, D., Silverman, E.K., Crapo, J.D.: Genetic epidemiology of COPD (COPDGene) study design. COPD 7(1), 32–43 (2010)

Author Index