

Capt Mark B. Skouson
AFIT/CI Program Completion Information
University of Illinois at Urbana-Champaign
July 2000

Abstract: This thesis presents a systematic study of deformable image transformations for nonrigidly aligning a template to an image. The study concentrates on an information-theoretic similarity measure, fluid deformation, and prior shape constraints.

The similarity between a target image and a template is measured based on their mutual information. Suitability of the mutual information measure for non-rigid-body image registration is systematically investigated. A mutual information bound is derived, and a gradient calculation, which scales linearly with the volume size, is presented. Modification of a fluid model proposed elsewhere is shown to retain the desirable properties of the deformation while allowing more efficient numerical implementation. Shape information is learned by performing eigenshape analysis on a training set of correct deformations of a single template to several typical segmentations. The most likely deformations are then promoted according to the learned shape information. The shape modeling technique does not require a prelabeling or ordering of points in the training set and can handle multiple shapes simultaneously. Based on these results, a new method is developed for nonrigidly aligning a template to a study image. The approach is robust to a wide variety of contrast variations and supports large, curved geometric variations.

This method has been experimentally validated using synthetic, magnetic resonance, and cryosection images. It is also incorporated into a brain image segmentation method under development at the University of Illinois. Potential applications include image segmentation, functional brain mapping, and automatic target recognition.

Author: Mark Brown Skouson

Title: Information-Theoretic Approach to Deformable Image
Transformations with Application to Brain Image
Segmentation

Military Rank: Captain

Service Branch: United States Air Force

Date: 2000

Number of Pages: 131

Degree Awarded: Doctor of Philosophy

Institution: University of Illinois at Urbana-Champaign

References

- [1] A. J. Abrantes and J. S. Marques. A class of constrained clustering algorithms for object boundary extraction. *IEEE Trans. Image Proc.*, 5(11):1507–1521, 1996.
- [2] J. Agris, R. deFigueiredo, G. Hillman, and T. Kent. A novel method for 3-D segmentation and volume estimation of brain compartments from MRI. In *Proc. Ann. Int. Conf. IEEE Engineering in Medicine and Biology Society*, volume 13, pages 68–70, 1991.
- [3] N. Ahuja. A transform for multiscale image segmentation by integrated edge and region detection. *IEEE Trans. Patt. Anal. Mach. Intell.*, 18:1211–1235, 1996.
- [4] R. Bajcsy and S. Kovacic. Multiresolution elastic matching. *Comp. Vis. Graph. Image Proc.*, 46:1–21, 1989.
- [5] E. Bardinet, L. D. Cohen, and N. Ayache. A parametric deformable model to fit unstructured 3D data. *Comp. Vision and Image Understanding*, 71(1):39–54, July 1998.
- [6] C. A. Bouman and M. Shapiro. A multiscale random field model for Bayesian image segmentation. *IEEE Trans. Image Proc.*, 3(2):166–177, 1994.
- [7] M. Bro-Nielsen and C. Gramkow. Fast fluid registration of medical images. In *Visualization in Biomedical Computing, 4th International Conference, VBC '96*, pages 267–276, 1996.
- [8] C. Broit. *Optimal Registration of Deformed Images*. PhD thesis, University of Pennsylvania, Philadelphia, PA, 1981.
- [9] M. S. Brown et al. Method for segmenting chest CT image data using an anatomical model: Preliminary results. *IEEE Trans. Med. Imaging*, 16(6):828–839, 1997.
- [10] R. C. Buck. *Advanced Calculus*. McGraw-Hill, St. Louis, MO, 1978.
- [11] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *Int. J. Comput. Vis.*, 22(1):61–79, 1997.
- [12] W. C. Chew. *Waves and Fields in Inhomogeneous Media*. IEEE Press, New York, 1995.
- [13] G. E. Christensen. *Deformable Shape Models for Anatomy*. PhD thesis, Washington University, Saint Louis, MO, 1994.
- [14] G. E. Christensen. Consistent linear-elastic transformations for image matching. In *LNCS*, volume 1613, pages 224–237, 1999.
- [15] G. E. Christensen, H. J. Johnson, J. W. Haller, J. Melloy, M. W. Vannier, and J. L. Marsh. Synthesizing average 3D anatomical shapes using deformable templates. In *SPIE Medical Imaging 1999: Image Processing*, volume 3661, pages 574–582, 1999.
- [16] G. E. Christensen, S. C. Joshi, and M. L. Miller. Volumetric transform of brain anatomy. *IEEE Trans. Med. Imaging*, 16(6):864–877, Jan. 1997.
- [17] G. E. Christensen, R. D. Rabbitt, and M. I. Miller. Deformable templates using large deformation kinematics. *IEEE Trans. Image Proc.*, 5(10):1435–1447, 1996.
- [18] L. D. Cohen. On active contour models and balloons. *CVGIP: Image Understanding*, 53(1):211–218, Mar. 1991.
- [19] A. Collignon et al. Automated multi-modality image registration based on information theory. In *Proc. Information Processing in Medical Imaging Conf.*, pages 263–274, 1995.
- [20] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam. Use of active shape models for locating structures in medical images. *Image and Vision Comp.*, 12(6):355–365, July/August 1994.
- [21] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models—their training and application. *Comp. Vision and Image Understanding*, 61(1):38–59, Jan. 1995.
- [22] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.
- [23] C. A. Davatzikos. Spatial transformation and registration of brain images using elastically deformable models. *Comp. Vision and Image Understanding*, 66(2):207–222, May 1997.
- [24] C. A. Davatzikos and J. L. Prince. An active contour model for mapping the cortex. *IEEE Trans. Med. Imaging*, 14:65–80, Mar. 1995.
- [25] N. Duta, A. K. Jain, and M.-P. Dubisson-Jolly. Learning 2D shape models. In *Comp. Vision Patt. Recog.*, volume 2, pages 8–14, 1999.

- [26] N. Duta and M. Sonka. Segmentation and interpretation of MR brain images: An improved active shape model. *IEEE Trans. Med. Imaging*, 17(6):1049–1062, Dec. 1998.
- [27] V. A. Epanechnikov. Nonparametric estimates of a multivariate probability density. *Theor. Probability Appl.*, 14:153–158, 1969.
- [28] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. Computers*, 22(1):67–92, 1973.
- [29] M. Frigo and S. G. Johnson. FFTW: An adaptive software architecture for the FFT. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98*, volume 3, pages 1381–1384, New York, NY, 1998. IEEE.
- [30] T. Gaens, F. Maes, D. Vandermuelen, and P. Suetens. Non-rigid multimodal image registration using mutual information. In *Medical Image computing and computer-assisted Intervention – MICCAI'98*, pages 1099–1106, 1998.
- [31] D. Geiger, A. Bupta, L. A. Costa, and J. Vlontzos. Dynamic programming for detecting, tracking, and matching deformable contours. *IEEE Trans. Patt. Anal. Mach. Intell.*, 17(3):294–302, Mar. 1995.
- [32] R. M. Gray. *Entropy and Information Theory*. Springer-Verlag, New York, 1990.
- [33] R. M. Gray and L. D. Davisson. *Random Processes, A Mathematical Approach for Engineers*. Prentice-Hall, Englewood Cliffs, NJ, 1986.
- [34] R. M. Haralick and L. G. Shapiro. Survey: Image segmentation techniques. *Comp. Vis. Graph. Image Proc.*, 29:100–32, 1985.
- [35] K. Haris, S. N. Efstratiadis, and N. Maglaveras. Watershed-based image segmentation with fast region merging. In *Int. Conf. Image Proc.*, volume 3, pages 338–342, Chicago, IL, 1998.
- [36] S. Harowitz and T. Pavlidis. Picture segmentation by a directed split-and-merge procedure. In *Proc. 2nd Int. Joint Conf. Pattern Recognition*, pages 424–433, 1974.
- [37] N. Hata, T. Dohi, S. Warfield, W. M. Wells, R. Kikinis, and F. A. Jolesz. Multimodality deformable registration of pre- and intraoperative images for MRI-guided brain surgery. In *Medical Image computing and computer-assisted Intervention – MICCAI'98*, pages 1067–1074, 1998.
- [38] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, Englewood Cliffs, NJ, 1999.
- [39] M. T. Heath. *Scientific Computing: An Introductory Survey*. McGraw-Hill, New York, 1997.
- [40] K. Held et al. Markov random field segmentation of brain MR images. *IEEE Trans. Med. Imaging*, 16(6):878–886, 1997.
- [41] A. K. Jain, Y. Zhong, and M.-P. Dubuisson-Jolly. Deformable template models: A review. *Sig. Proc.*, 71(2):109–129, 1998.
- [42] X. Ji, S. Wang, M. B. Skouson, and Z.-P. Liang. Brain image segmentation by multiscale analysis and template matching. In *Int. Soc. Magn. Reson. Med.*, page 2186, Philadelphia, PA, 1999.
- [43] R. A. Johnson and D. W. Wichern. *Multivariate Statistics, A Practical Approach*. Chapman and Hall, London/New York, 1988.
- [44] S. C. Joshi. *Large Deformation Diffeomorphisms and Gaussina Random Fields for Statistical Characterization of Brain Sub-manifolds*. PhD thesis, Washington University, Saint Louis, MO, 1998.
- [45] T. Kanungo, B. Dom, W. Niblack, and D. Steele. A fast algorithm for MDL-based multi-band image segmentation. In *Comp. Vision Patt. Recog.*, pages 609–616, 1994.
- [46] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *Int. J. Comput. Vis.*, 1(4):321–331, 1987.
- [47] B. B. Kimia, A. R. Tannenbaum, and S. W. Zucker. Shapes, shocks, and deformations I: The components of two-dimensional shape and the reaction diffusion space. *Int. J. Comput. Vis.*, 15:189–224, 1995.

- [48] R. Kimmel, N. Kiryati, and A. M. Bruckstein. Analysing and synthesizing images by evolving curves with the Osher-Sethian method. *Int. J. Comput. Vis.*, 24(1):37–55, 1997.
- [49] Y. G. Leclerc. Constructing simple stable descriptions for image partitioning. *Int. J. Comput. Vis.*, 3:73–102, 1989.
- [50] S. Lobregt and M. A. Viergever. A discrete dynamic contour model. *IEEE Trans. Med. Imaging*, 14(1):12–24, Mar. 1995.
- [51] F. Maes. *Segmentation and Registration of Multimodal Medical Images*. PhD thesis, Katholieke Universiteit Leuven, Leuven, Belgium, 1998.
- [52] F. Maes, A. Collignon, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Trans. Med. Imaging*, 16:187–198, 1997.
- [53] J. B. A. Maintz, E. H. W. Meijering, and M. A. Viergever. General multimodal elastic registration based on mutual information. In *SPIE Medical Imaging 1998: Image Processing*, volume 3338, pages 144–154, San Diego, CA, 1998.
- [54] R. Malladi, J. A. Sethian, and B. C. Vemuri. Shape modeling with front propagation: A level set approach. *IEEE Trans. Patt. Anal. Mach. Intell.*, 17(2):158–175, 1995.
- [55] S. Mallat and W. L. Hwang. Singularity detection and processing with wavelets. *IEEE Trans. Inf. Theory*, 38:617–643, 1992.
- [56] D. Marr and E. C. Hildreth. Theory of edge detection. In *Proc. Roy. Soc. London, B*, volume 207, pages 181–217, 1980.
- [57] N. J. I. Mars et al. Time delay estimation in non-linear systems using average amount of mutual information analysis. *Signal Processing*, 4:139–153, 1982.
- [58] T. McInerney and R. Kikinis. An object-based volumetric deformable atlas for the improved localization of neuroanatomy in MR images. In *Medical Image computing and computer-assisted Intervention – MICCAI’98*, pages 861–869, 1998.
- [59] T. McInerney and D. Terzopoulos. Topologically adaptable snakes. In *Proc. Fifth Int. Conf. on Computer Vision (ICCV’95)*, pages 840–845, Cambridge, MA, June 1995.
- [60] T. McInerney and D. Terzopoulos. Deformable models in medical image analysis: A survey. *Medical Image Analysis*, 1(2):91–108, 1996.
- [61] R. Moddemeyer. On estimation of entropy and mutual information on continuous distributions. *Signal Processing*, 16(3):233–248, Mar. 1989.
- [62] W. J. Niessen et al. Geodesic deformable models for medical image analysis. *IEEE Trans. Med. Imaging*, 17(4):634–641, Aug. 1998.
- [63] W. J. Niessen, K. L. Vincken, J. Weickert, and M. A. Viergever. Three-dimensional MR brain segmentation. In *Sixth International Conf. on Computer Vision. Bombay, India*, pages 53–58, Jan. 1998.
- [64] S. Osher and J. A. Sethian. Fronts propagating with curvature dependent speed: Algorithms based on Hamiltonian-Jacobi formulation. *J. of Computational Physics*, 79:12–49, 1988.
- [65] T. Poggio, V. Torre, and C. Koch. Computational vision and regularization theory. *Nature*, 317(6035):314–319, 1985.
- [66] S. P. Raya. Low-level segmentation of 3D magnetic resonance brain images—a rule-based system. *IEEE Trans. Med. Imaging*, 9:327–337, 1990.
- [67] M. Rosenblatt. Curve estimates. *Annals of Mathematical Statistics*, 42(6):1815–1842, 1971.
- [68] A. P. Sage and I. W. Chelsea C. *Optimum Systems Control*. Prentice Hall, Englewood Cliffs, NJ, 1977.
- [69] Y. Salu and J. Tilton. Classification of multispectral image data by binary diamond neural network and by nonparametric, pixel-by-pixel methods. *IEEE Tran. on Geoscience and Remote Sensing*, 31(3):606–616, 1993.
- [70] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.
- [71] M. B. Skouson and M. E. Kowalski. A simple and efficient method for image segmentation with deformable templates. In *IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 31–35, Austin, Texas, Apr. 2000.

- [72] M. B. Skouson and Z.-P. Liang. Template deformation by maximizing mutual information. In *Proc. BMES/EMBS*, page 1162, Atlanta, GA, 1999.
- [73] M. Sonka et al. Knowledge-based interpretation of MR brain images. *IEEE Trans. Med. Imaging*, 15(4):443–452, 1998.
- [74] H. Stark and J. W. Woods. *Probability, Random Processes, and Estimation Theory for Engineers*. Prentice Hall, Englewood Cliffs, NJ, 1994.
- [75] C. Studholme, D. L. G. Hill, and D. J. Hawkes. Incorporating connected region labelling into automated image registration using mutual information. In *Proc. MMBIA*, pages 23–31, 1996.
- [76] B. Szekely, A. Kelemen, C. Brechbuhler, and G. Gerig. Segmentation of 3D objects from MRI volume data using constrained elastic deformations of flexible Fourier surface models. In *CVRMed'95*, pages 495–505, Berlin, Germany, 1995.
- [77] M. Tabb and N. Ahuja. Multiscale image segmentation by integrated edge and region detection. *IEEE Trans. Image Proc.*, 6(5):642–654, 1997.
- [78] D. Terzopoulos and D. Metaxas. Dynamic 3-D models with local and global deformations: Deformable superquadrics. *IEEE Trans. Patt. Anal. Mach. Intell.*, 13(7):703–714, July 1991.
- [79] B. C. Vemuri and A. Radisavljevic. Multiresolution stochastic hybrid shape models with fractal priors. *ACM Transactions on Graphics*, 13(2):177–120, Apr. 1994.
- [80] K. L. Vincken, A. S. E. Koster, and M. A. Viergever. Probabilistic multiscale image segmentation. *IEEE Trans. Patt. Anal. Mach. Intell.*, 19:109–120, 1997.
- [81] P. Viola and W. W. Wells. Alignment by maximization of mutual information. In *Proc. 5th Int. Conf. Computer Vision*, pages 15–23, 1995.
- [82] P. A. Viola. *Alignment by Maximization of Mutual Information*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1995.
- [83] E. W. Weisstein. *CRC Concise Encyclopedia of Mathematics*. CRC Press, Boca Raton, FL, 1999.
- [84] W. M. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, 1(1):25–51, 1996.
- [85] J. West et al. Comparison and evaluation of retrospective intermodality brain image registration techniques. *J. Comput. Assist. Tomog.*, 21(4):554–566, 1997.
- [86] B. Widrow. The rubber mask technique. *Patt. Recog.*, 5(3):175–211, 1973.
- [87] C. Xu and J. L. Prince. Snakes, shapes, and gradient vector flow. *IEEE Trans. Image Proc.*, 7(3):359–369, 1998.
- [88] A. Yezzi, S. Kichenassamy, A. Kumar, P. Oliver, and A. Tannenbaum. A geometric snake model for segmentation of medical imagery. *IEEE Trans. Med. Imaging*, 16:199–209, 1997.
- [89] A. P. Zijdenbos and B. B. Dawant. Brain segmentation and white matter lesion detection in MR images. *Critical Reviews in Biomedical Engineering*, 22(5/6):401–465, 1994.
- [90] I. G. Zubal, C. R. Harrell, E. O. Smith, Z. Rattner, G. R. Gindi, and P. B. Hoffer. Computerized three-dimensional segmented human anatomy. *Medical Physics*, 21(2):199–302, February 1994.
- [91] S. W. Zucker. Survey region growing: Childhood and adolescence. *Comp. Graph. and Image Proc.*, 5:382–399, 1976.

© Copyright by Mark Brown Skouson, 2000

INFORMATION-THEORETIC APPROACH TO DEFORMABLE IMAGE
TRANSFORMATIONS WITH APPLICATION TO BRAIN IMAGE
SEGMENTATION

BY

MARK BROWN SKOUSON

B.S., Brigham Young University, 1993

M.S., Brigham Young University, 1994

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2000

Urbana, Illinois

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
THE GRADUATE COLLEGE

JUNE 2000
(date)

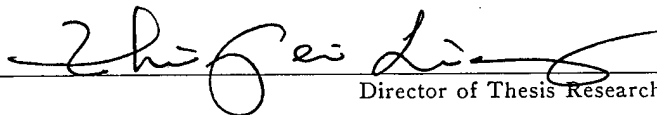
WE HEREBY RECOMMEND THAT THE THESIS BY

MARK BROWN SKOUSON

ENTITLED INFORMATION-THEORETIC APPROACH TO DEFORMABLE
IMAGE TRANSFORMATIONS WITH APPLICATION TO BRAIN
IMAGE SEGMENTATION

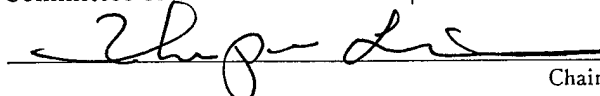
BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

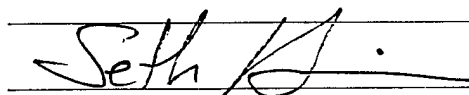
THE DEGREE OF DOCTOR OF PHILOSOPHY

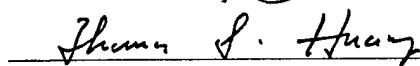

Director of Thesis Research

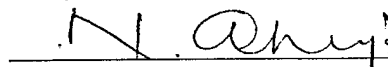
Head of Department

Committee on Final Examination†


Chairperson







† Required for doctor's degree but not for master's.

ABSTRACT

This thesis presents a systematic study of deformable image transformations for non-rigidly aligning a template to an image. The study concentrates on an information-theoretic similarity measure, fluid deformation, and prior shape constraints.

The similarity between a target image and a template is measured based on their mutual information. Suitability of the mutual information measure for non-rigid-body image registration is systematically investigated. A mutual information bound is derived, and a gradient calculation, which scales linearly with the volume size, is presented. Modification of a fluid model proposed elsewhere is shown to retain the desirable properties of the deformation while allowing more efficient numerical implementation. Shape information is learned by performing eigenshape analysis on a training set of correct deformations of a single template to several typical segmentations. The most likely deformations are then promoted according to the learned shape information. The shape modeling technique does not require a prelabeling or ordering of points in the training set and can handle multiple shapes simultaneously. Based on these results, a new method is developed for nonrigidly aligning a template to a study image. The approach is robust to a wide variety of contrast variations and supports large, curved geometric variations.

This method has been experimentally validated using synthetic, magnetic resonance, and cryosection images. It is also incorporated into a brain image segmentation method under development at the University of Illinois. Potential applications include image segmentation, functional brain mapping, and automatic target recognition.

ACKNOWLEDGMENTS

First of all, I would like to acknowledge and thank my advisor, Professor Zhi-Pei Liang, for his support, corrections, suggestions, example, training, and enthusiasm. He always placed my progress above his own.

I thank Professors Narendra Ahuja, Thomas Huang, and Seth Hutchinson for agreeing to serve on my committee and for their helpful comments. I also thank Professors Joseph Smith, Arne Pearlstein, and Michael Olsen for helping me understand a little about fluid mechanics.

I also wish to recognize the contributions my fellow classmates have made to this thesis: Hao Pan, Xiuquan (Jim) Ji, Song Wang, Chris Hess, Brian Stewart, Mark Kowalski, Greg Toussaint, Ananya Sen Gupta, and Quji (Alan) Guo. We had many hours of discussions which helped me revise or solidify my positions on various concepts. Jim Ji also wrote the first code I used for the fluid deformation. Song Wang helped formulate the statistical combination of the deformed template result and the multiscale region analysis. Marc Kowalski helped immensely with the modifications to the fluid transform and quantifying the computational speedup. Alan Guo helped me work through the mutual information bound. Others who contributed include those who provided medical images; the FFTW group, for their computer code; Professor Ahuja's group, for the multiscale region analysis code; and those whose research I built upon.

I also appreciate the support I received from my wife, Laura; my children, Dallin and Kate; and my parents. Their faith and encouragement helped me to persevere. And of course, I give final credit to the hand of Providence.

TABLE OF CONTENTS

CHAPTER		PAGE
1	INTRODUCTION	1
1.1	Motivation	2
1.2	Problem Formulation	4
1.3	Overview of the Proposed Method	6
1.4	Main Contributions	7
1.5	Organization of the Thesis	8
2	BRIEF LITERATURE REVIEW	9
2.1	Active Contours	9
2.2	Implicit Snakes	10
2.3	Deformable Templates	11
2.4	Information-Based Template Matching	12
2.5	MDL-Based Image Partitioning	13
2.6	Deformations Using Shape Information	14
3	RELEVANT TOPICS FROM PROBABILITY AND INFORMATION THEORY	16
3.1	Random Variables	16
3.1.1	Probability density functions and probability mass functions . . .	17
3.1.2	The Gibbs distribution	18
3.1.3	Functions of random variables	19
3.1.4	Joint and conditional probabilities	20
3.1.5	Simple statistics of random variables	21
3.1.6	Central limit theorem	22
3.2	Parzen Window Density Estimation	23
3.3	Information	31
3.3.1	Entropy	31
3.3.2	Mutual information	32
3.3.3	Kolmogorov complexity	33
4	TEMPLATE ALIGNMENT AND BODY FORCE	34
4.1	Alignment Based on Mutual Information	34
4.1.1	Squared-error and correlation alignment metrics	36

4.1.2	Mutual information for alignment	40
4.1.3	Parzen windowing for MI	42
4.2	Using the MI to Find a Body Force	44
4.3	MI Bias and Variance	48
4.3.1	Bias	49
4.3.2	Variance	49
4.3.3	Generalization for Parzen windows	51
4.4	Bound on the Maximal Mutual Information	57
4.4.1	Discussion of the bound	59
4.5	Optimality of Mutual Information	62
4.5.1	Maximum likelihood optimality	64
4.6	Summary	66
5	A FLUID-BASED TRANSFORM MODEL	68
5.1	Terminology and Relevant Theorems	68
5.1.1	Topology conservation	68
5.1.2	Large, nonlinear deformations	69
5.1.3	Eulerian formulation	70
5.1.4	Pertinent theorems	70
5.2	Fluid Model	72
5.3	Discrete Approximation to the Fluid Model	73
5.4	Modified Fluid Model	74
5.5	Fluid Model as a Prior Model	78
5.6	Implementation Issues	78
5.7	Experimental Validation	79
5.8	Summary	89
6	INCLUSION OF SHAPE INFORMATION	90
6.1	The Point Distribution Model	91
6.2	The Shape Model	93
6.2.1	Model statistics from training samples	94
6.3	Inclusion of Shape Information into the Deformable Model	95
6.4	Implementation Issues	97
6.5	Experimental Results	98
6.6	Summary	99
7	A HYBRID BRAIN IMAGE SEGMENTATION METHOD	102
7.1	Brain Image Segmentation	102
7.1.1	Feature extraction: Multiscale region segmentation	103
7.1.2	Geometric constraints	106
7.1.3	Final classification	108
7.2	Implementation Issues	110
7.3	Experimental Results	111
7.4	Summary	115

8 CONCLUSION	116
8.1 Areas of Further Research	117
8.1.1 Incorporating other types of learning	117
8.1.2 Better integration of deformation and region information	118
8.1.3 Hierarchical deformation	118
8.1.4 Other data sets	118
APPENDIX A IDENTITIES IN INFORMATION THEORY	120
APPENDIX B DERIVATION OF THE FLUID MODEL COST FUNCTIONAL	122
REFERENCES	124
VITA	131

LIST OF TABLES

Table	Page
4.1 Measurements of the utility of (4.21) when Parzen windowing is used. Comparison is made between the MI gradient calculation using (4.21) and the exact calculation. Maximum and average magnitude differences relative to the maximum body force magnitude are shown.	47

LIST OF FIGURES

Figure	Page
1.1 An example of template matching.	2
1.2 Several 2-D brain image examples.	4
1.3 Schematic of the proposed deformation procedure.	7
3.1 Plot of the PDF from which samples were drawn for tests corresponding to Figures 3.2, 3.3, and 3.4.	28
3.2 Plot of the error between the estimated and actual PDF as a function of Parzen window width b_N for different window shapes.	29
3.3 Error between the estimated and actual PDFs as a function of window width for a Gaussian window and 500 samples.	29
3.4 Error between the estimated and actual PDFs as a function of window width for a Gaussian window and 5000 samples.	30
3.5 Plots of the actual and estimated distributions using a Gaussian Parzen window with several widths.	30
4.1 A general imaging model.	35
4.2 A one-dimensional structure and the resulting image based on simply adding noise.	37
4.3 Alignment error for all possible shifts of the structure in Figure 4.2. . . .	38
4.4 A one-dimensional structure and the resulting image based on the imaging function of (4.9).	38
4.5 Alignment error for all possible shifts of the structure in Figure 4.4. . . .	39
4.6 Alignment error for the structure in Figure 4.4 when the imaging function was erroneously assumed to be the identity function.	39
4.7 Mutual information alignment error curve for the example structure and image shown in in Figure 4.2.	41
4.8 Mutual information alignment error curve for the structure and image shown in Figure 4.4.	42
4.9 Visual explanation of the partial volume effect.	43
4.10 1-D image example showing the utility of Parzen window probability estimation in the presence of partial volume effects.	44

4.11 Plot of the MI between images (a) and (b) as (b) was scaled smaller and larger.	47
4.12 Plot of the MI between images (a) and (b) as (b) was scaled smaller and larger.	47
4.13 Plot demonstrating predictability of N -bias.	53
4.14 Demonstration of the predictability of N -bias and S -bias when Parzen windowing is used.	54
4.15 Demonstration of the predictability of entropy estimation variance. . . .	55
4.16 Demonstration of the predictability of entropy estimation variance when Parzen windowing is used.	55
4.17 Aligning error curves using MI and Gaussian-shaped Parzen windows of varying widths.	56
4.18 Normalized alignment error curves using MI and Gaussian-shaped Parzen windows of varying widths.	56
4.19 Synthetic image with multiplicative noise and several possible segmentations, including one that achieves the bound.	60
4.20 Real cryosection image and several possible segmentations, including one that achieves the bound.	61
4.21 Bound achieving template separated by graylevel.	61
4.22 Relationship between $\log(G(\mathbf{a}))$, the derived bound, and $I(\mathbf{d}; \tilde{\mathbf{a}})$, where $\mathbf{u}$ is a scaling transform.	62
4.23 Symmetry problem example.	63
4.24 A many-to-one graylevel mapping problem example.	64
5.1 Homeomorphic vs. nonhomeomorphic.	69
5.2 Illustration of the relationship between a point fixed point in the deformed template and points in the original template as time varies.	71
5.3 Demonstration of orthogonally induced displacements.	75
5.4 A second demonstration of orthogonally induced displacements.	75
5.5 A third demonstration of orthogonally induced displacements.	79
5.6 Comparison between the original and modified fluid models.	80
5.7 Illustration of the fluid transform's ability to accommodate large, curved deformations.	80
5.8 Average time required to solve for the instantaneous velocity field in the course of the patch to "C" experiment at various image sizes.	81
5.9 Peak time required to solve for the instantaneous velocity field in the course of the patch to "C" experiment at various image sizes.	82
5.10 Three-dimensional patch to "C" experiment.	83
5.11 Template deformation to an MR-PD medical image.	84
5.12 Deformation to MR-T1 and cryosection medical images.	84
5.13 Deformation to several images using different Parzen window widths when calculating the body force.	86
5.14 Deformation to several images using various values for μ_2 in (5.9).	87

5.15	Deformation to several other images using various values of μ_2	88
6.1	Demonstration of the effectiveness of shape information in the deformation process.	100
6.2	Demonstration of the effectiveness of shape information in the deformation process (continued).	101
7.1	Flowchart of the described brain image segmentation method.	103
7.2	Multiscale analysis performed on an image.	106
7.3	Extracting the brain silhouette.	112
7.4	Comparison between segmentation based on template deformation and the combination of template deformation and multiscale region analysis. . . .	113
7.5	Comparison between segmentation based on template deformation and the combination of template deformation and multiscale region analysis (continued).	114

LIST OF SYMBOLS AND ABBREVIATIONS

$\tilde{A}$	Random variable associated with the template
$\mathbf{a}$	Template
$\tilde{a}$	A graylevel of the template
$\tilde{a}(i)$	An element of the template
$\tilde{\mathbf{a}}$	Deformed template
ASD	Allowable shape domain
B	Bound on mutual information
$\mathbf{b}$	Body force
$\mathbf{b}_d$	Data-dependent body force
$\mathbf{b}_p$	Shape-dependent body force
C	Imaging function
CDF	Cumulative distribution function
CSF	Cerebral spinal fluid
D	Random variable associated with the study image
$\mathbf{d}$	Study image data data
d	A graylevel of the study image
$d(i)$	An element of the study image
E	Expectation function
H	Entropy
$\mathbf{h}$	coordinate transform
I	Information or mutual information
$\mathcal{L}$	Internal model constraining deformation
$K_{\cdot,\cdot}$	Histogram random variable
$k_{\cdot,\cdot}$	Histogram variable
M	Variable representing a generic image alignment metric
MI	Mutual information
N_d	Number of graylevels in the study image
N_a	Number of graylevels in the template
N_x	Number of pixels along the vertical side of the image
N_y	Number of pixels along the horizontal side of the image
N_p	Number of pixels in the image

P	Probability
PDF	Probability density function
PDM	Point distribution model
PMF	Probability mass function
R, R_1, R_2, R_b	Parzen windows
r.v.	Random variable
s	Segmentation or structure of underlying object being imaged
$\hat{s}$	Estimate of the structure of the study image
t	Time
$\mathbb{U}$	Set of all possible displacement fields
$\mathbf{u}$	Displacement field
$\mathbf{u}^*$	Optimal displacement field
$\mathbf{v}$	Velocity of the displacement field
VAR	Variance function
$\mathbb{X}$	Set of all possible position vectors
$\mathbf{x}$	Position vector for the image and template
x, y, z	Element of the position vector
$\hat{x}_i : i = 1, 2$	Orthogonal unit vectors for 2-D Cartesian space
η	Noise
μ_1, μ_2, μ_3	Parameters determining characteristics of the fluid model
μ_4	Parameter determining emphasis on shape information
$\nabla_{\mathbf{u}}$	Directional gradient with respect to $\mathbf{u}$
ϕ	Exogenous factors
σ	Multiscale parameter, or variance of a normal distribution

CHAPTER 1

INTRODUCTION

It is a common practice in pattern recognition problems to match a candidate pattern to one of perhaps several prototype patterns. For all but the simplest applications, maintaining a prototype database with all possible variations of an input pattern is impractical. Early research concentrated on allowing several simple transformations of a prototype pattern to alleviate the need for so many prototypical patterns. As research has matured in this area, allowable transformations have become more sophisticated, even to include a large number of nonrigid deformations [1, 2].

The main concepts of deformable models have origins dating back to at least the early 1970s. For example, the works of Widrow [3] and Fischler and Elschlager [4] involved rubber masks or rigid components held together by “springs” for use in elastically matching a model to an image being studied. Interest in deformable models has increased in recent years due largely to two factors. (a) Deformable models are an effective way to combine low-level image information with higher-level prior information. (b) Increases in computational speeds of computers have enabled the creation of deformable models with sufficient sophistication to begin to handle practical problems. Figure 1.1 illustrates the potential use of a deformable matching technique; a single template may be used to describe a broader class of candidates than would a rigid technique.

However, the area of deformable models is not mature. Existing limitations prevent current deformable matching techniques from solving many practical computer vision

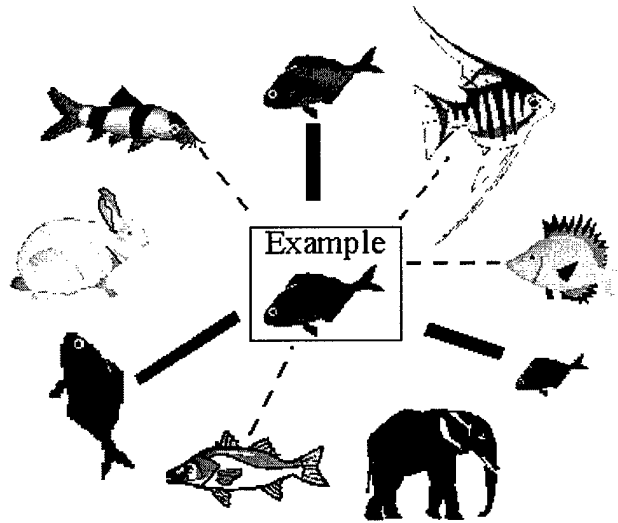


Figure 1.1 An example of template matching. The example fish matches only three fish (solid lines) using rigid transformations. Deformable transformations may allow the example to match other fish (dashed lines). (Adapted from [1].)

problems. For example, many techniques rely on absolute intensity information, precluding use in an unknown contrast environment. Others rely on intensity gradient information and ignore information from within region boundaries. Such methods are sensitive to noise and initialization. Additionally, many methods are formulated for single object deformation, so implementation in the presence of multiple contours, objects, etc., is problematic. Finally, in most cases, deformations are based on prior information from only a small number of localized points. Global “shape” information is largely ignored, so deformations may occur which are not consistent with prior knowledge of the system. This dissertation describes a template deformation method which presents significant advances in overcoming these limitations.

1.1 Motivation

The need for more automated image processing methods is pressing. With the rapid advances in imaging technology, it is now routine to acquire large image data sets easily and quickly; in fact, the ability to acquire images has thus far surpassed our ability to

efficiently analyze and interpret them. This is particularly true in the area of medical imaging.

Magnetic resonance imaging (MRI), x-ray computed tomography (CT), and ultrasonic imaging are quite commonplace, while current clinical segmentation techniques are, at best, semiautomated. Magnetic resonance (MR) images are routinely used to help diagnose disorders of the brain; however, the manual segmentation associated with the diagnosis can be quite time-consuming and prone to error, bias, and nonrepeatability.

For example, the progression of some illnesses, such as AIDS, can be determined by measuring changes in the volume of white matter, gray matter, or the ratio between the two [5]. Manual-based measurements are potentially biased because measurements are taken over a period of time by perhaps different experts. An automated segmentation method could reduce these biases. In other medical applications it is important to localize key anatomical regions and structures. These structures can then be used to help neuroscientists examine the physiological processes in functional studies. An automated means of localizing and labeling these structures would certainly benefit researchers in this area. The same information would also be useful to surgeons to aid in surgical planning.

Figure 1.2 shows several medical images which could be considered to belong to a single class because they were acquired from the “same” anatomical location of several human brains. As can be seen from the image, some portions of the anatomy have a highly regular shape — the scalp and skull for example — while the inner structures (white matter and gray matter) are highly irregular and can exhibit large, curved differences. Additionally, depending on the imaging process, the average graylevel intensities of corresponding structures in the different images (i.e., the image contrast) differ widely. These issues present problems for existing deformable template methods.

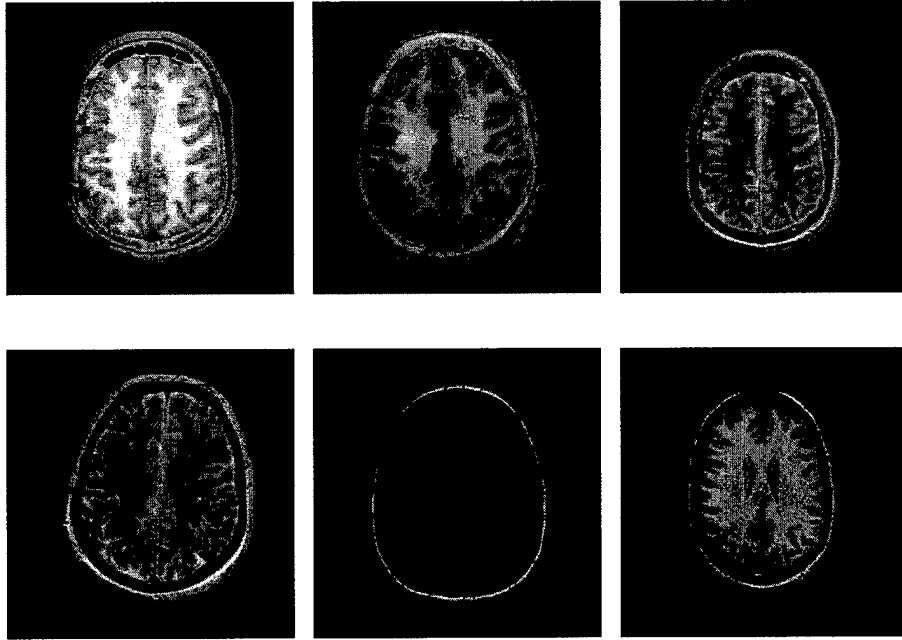


Figure 1.2 Several 2-D brain image examples. Images are axial views of the supraventricular region of different human brains using different imaging parameters.

1.2 Problem Formulation

Consider a fixed image being studied $\mathbf{d}(\mathbf{x})$ with an underlying structure $\mathbf{s}(\mathbf{x})$, where $\mathbf{x}$ represents the domain on which both $\mathbf{d}$ and $\mathbf{s}$ are defined. An imaging process can be modeled as

$$\mathbf{d}(\mathbf{x}) = C(\mathbf{s}(\mathbf{x}), \phi) + \eta \quad (1.1)$$

The actual intensity values in $\mathbf{d}$ are determined by the underlying structure $\mathbf{s}$, the imaging function $C(\cdot, \cdot)$, exogenous factors ϕ , and noise η . The exogenous factors are factors such as lighting, temperature, etc., that are apart from the actual object being imaged, but that affect the intensities in the image.

The template deformation problem can be defined as follows: Given an image $\mathbf{d}(\mathbf{x})$ and a template $\mathbf{a}(\mathbf{x})$ find a coordinate displacement field $\mathbf{u}^*$ from a set of possible displacement fields $\mathbb{U}$ such that when it is applied to $\mathbf{a}$, $\mathbf{a}(\mathbf{x} - \mathbf{u}^*(\mathbf{x}))$ is most closely aligned

with $\mathbf{d}$. More concisely,

$$\mathbf{u}^* = \arg \max_{\mathbf{u} \in \mathbf{U}} M(\mathbf{d}, \mathbf{a}(\mathbf{u})) \quad (1.2)$$

where M is a metric which measures the match between the template and study.

The problem is often formulated in terms of forces acting on the template. The force balance equation is

$$\mathcal{L}\mathbf{u} + \mathbf{b} = 0 \quad (1.3)$$

Using this formulation, the template deformation is dependent on an external body force $\mathbf{b}$, which is a function of M , and an internal model $\mathcal{L}$. The internal model maintains allowable deformations of the template, and the body force drives the template into alignment with the study image data. The “correct” deformation is assumed when the balance of forces is met. Different deformable models utilize different models for $\mathcal{L}$ and $\mathbf{b}$.

A desirable property of an internal model is that it allow deformations which are likely, given prior knowledge about the imaged object, while rejecting deformations that contradict the prior knowledge. However, as is illustrated in Figure 1.2, not all image regions will exhibit the same degree of variability. Variations in the contour of the skull are relatively small compared to the size of the skull, but the white matter variations can be on the same order of magnitude as the white matter region itself. In addition to the large variations present in the white matter, many of the finger-like regions are curved, so $\mathcal{L}$ should be sufficiently general for $\mathbf{u}$ to follow large, curved (nonlinear) trajectories.

As the internal model becomes general enough to accommodate large, nonlinear variations, the dimensionality of the deformation space increases, thus increasing the possibility for over-deformation. Rigid-body transformations have few degrees of freedom — on the order of 10. For non-rigid-body deformations the degrees of freedom are much higher — on the order of 100,000 for a 256×256 2-D image. This greatly increases both the complexity of the search space and the likelihood becoming trapped in a local extremum. In the rigid-body case there are many more data than degrees of freedom, so

redundancies exist. In the non-rigid-body case, the data and degrees of freedom are on the same order of magnitude, so the deformation result is much more sensitive to noise and other uncertainties in the data. A fundamental challenge, then, is to define a model that is sufficiently general to allow likely large, nonlinear deformations, but not allow unacceptable over-deformations.

A second challenge is defining a body force independent of the image contrast. As also illustrated in Figure 1.2, the contrast of the images may vary within an image class, so body forces relying on metrics such as squared intensity differences are inadequate. Some methods avoid using absolute intensity information by using the image intensity gradient instead. As has been previously mentioned, these methods ignore information within regions and are sensitive to noise and initialization.

A third challenge is deformation in the presence of multiple contours or objects. Many deformable methods that are well defined for a single object or contour are not easily extendible to multiple contours.

1.3 Overview of the Proposed Method

This thesis describes a new template deformation method with significant developments in overcoming the challenges just described. The model is made contrast-tolerant by computing a body force based on the mutual information between the template and study images. A viscous fluid internal model is used to allow large, curved deformations while maintaining the topology of the template. Over-deformations controlled through an additional body force based on prior shape information. Due to the volumetric nature of the fluid model, deformation of multiple objects is inherent. More explicitly, (1.3) is rewritten as

$$\mathcal{L}\mathbf{u} + \mathbf{b}_d + \mathbf{b}_p = 0 \quad (1.4)$$

where $\mathcal{L}$ is a fluid transformation operator, $\mathbf{b}_d$ is a force based on the mutual information between the image data and the template, and $\mathbf{b}_p$ is a force based on learned shape information.

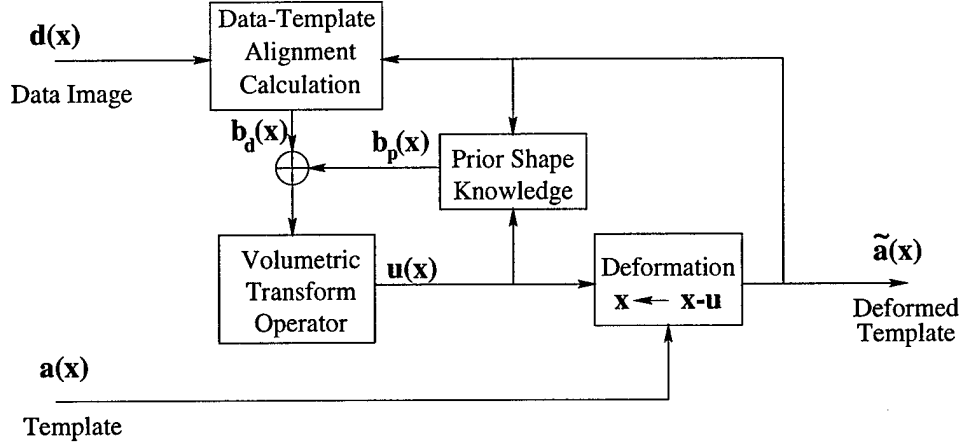


Figure 1.3 Schematic of the proposed deformation procedure. A template is incrementally deformed to match the image data and the prior knowledge.

Figure 1.3 shows a pictorial schematic of the method. A study image $\mathbf{d}$ and a template $\mathbf{a}$ are input into the system. The template is iteratively deformed according to (1.4).

The work herein concentrates on 2-D input images, although some 3-D data was used in the generation of experimental results.

1.4 Main Contributions

- *Contrast-independent template alignment.* An in-depth study of mutual information as it applies to nonrigid template deformation is presented. Conditions for optimality are defined. A study of the bias and variance of the mutual information between images is also presented. Limitations on the metric are investigated.
- *Mutual information bound.* A bound on the maximal mutual information is derived. Study of the bound helps to identify limitations of MI as an alignment metric.
- *Body force based on maximization of mutual information.* A formula is derived by which the mutual information between a study image and template is converted to a body force. The complexity of the body force calculation is linearly proportional to the image size.

- *Fluid deformation model.* A fluid model is developed. The model maintains desirable properties of an existing fluid model while being computationally more efficient. The model is proven to maintain the topology of the template throughout deformation.
- *Prior shape information.* A model is developed to capture both inter- and intrashape variations among regions of the template. The model is demonstrated to improve the deformation result.
- *Application to brain image segmentation.* The developed model is incorporated into a brain image segmentation method [6] that is part of an ongoing project at the Beckman Institute for Advanced Science and Technology on the University of Illinois at Urbana-Champaign.

1.5 Organization of the Thesis

The next chapter gives a brief review of research related to the work herein. Chapter 3 summarizes relevant concepts from probability and information theory. The theory and description of the proposed deformable model are found in Chapters 4, 5, and 6, which discuss the information-theoretic body force, the fluid transform, and shape-based constraints, respectively. Application to brain image segmentation is shown in Chapter 7, and Chapter 8 gives concluding remarks and discusses areas for further research.

CHAPTER 2

BRIEF LITERATURE REVIEW

2.1 Active Contours

As mentioned previously, the main concepts of deformable models have origins dating back to at least the early 1970s. Work by Widrow [3] involved rubber masks, and Fischler and Elschlager [4] used rigid components held together by “springs” to elastically match a model to a test image.

Deformable methods did not become widely popular, however, until Kass et al. [7] reported the development of *active contours*, or *snakes*. The snake model consists of an elastic contour that is deformed according to a cost function. The contour is active in the sense that it can adapt itself to fit the image. The cost function is the summation of an image-dependent term, usually based on the image intensity gradient, and an image-independent regularization term. The cost function is minimized using gradient descent. Essentially, an operator specifies an initial contour near a region boundary and the method actively seeks a close edge within the image under regularizing constraints of tension and rigidity. The contour is considered to be fully deformed when the regularization forces acting on the contour balance out the attractive forces of the image.

Since its introduction, several improvements and modifications to the active contour method have been proposed including modifications on the external force [8, 9], internal force [10], and the addition of a balloon force to improve performance in noisy situations [11]. Others have investigated different optimization strategies [12], use of snakes in

analysis of multiobject scenes [13], and the relationship of snakes to other algorithms [14].

The extension of snakes to multiobject scenes is of major importance to brain image segmentation because of the multiple structures and tissue classes that exist. A straightforward approach to the multiple surface deformation is to deform each surface separately using snakelike methods and then interpolate between the deformed surfaces using an elastic warping [15]. Because the initial surfaces are deformed independently, the spatial relationship between the contours is ignored as they deform.

2.2 Implicit Snakes

Another method that has been investigated for segmentation of multiple objects is implicit snakes or propagating fronts [16, 17, 18, 19, 20, 21]. Implicit snakes are based on the level-set theory introduced by Osher and Sethian [22]. In this approach the multiple contours (or surfaces) to be deformed are represented by a level set of a higher-dimensional function. The contours are then indirectly deformed based on the evolution of the higher-dimensional function, which is based on edge attraction similar to the snake method. The implicit snake method is analogous to a grass-fire propagation. The flame front propagates through a grass field (the image), slows when the image gradient is high, and finally stops at the edge. Also, once a particle (pixel) has been “burnt” (crossed by the front) it stays burnt, so the contour is free to split or merge. Use of implicit snakes for multiple objects is limited, though, since the original formulation does not provide for concentric contours. Extension to handle concentric contours has had limited success. In [17], Niessen et al. found tracking multiple level-sets to be incompatible with balloon forces used to overcome insignificant edges. An additional problem is that the implicit snake model places no restriction on the topology of the structure being segmented. This can easily lead to an over-segmentation of the image, resulting in the need to merge regions. The implicit snake model also inherits some of the problems of the original snake model: dependency on only image edges and sensitivity to noise. Finally,

Niessen et al. [17] concluded, after some investigation of the usefulness of implicit snakes in medical imaging applications, that without user-supplied initialization, topological constraints, and *a priori* knowledge, it is highly unlikely that, for example, in an MR brain image a gradient-dependent metric will suffice for delineation of the white matter surface. Therefore, implicit snake models do not seem to be suited for some of the more challenging problems in medical image segmentation.

2.3 Deformable Templates

An alternative approach to multiple surface deformation is template (atlas) deformation [23, 24, 25, 26]. Template deformations model the entire image volume as a 3-D lattice such that each element in the lattice is connected to each of its neighbors. A force on any piece of the volume is distributed to other elements in the volume through the neighbor-to-neighbor connections. Deformable template methods have the advantage over snakelike methods in that spatial relationships among volume elements are calculated automatically through the interconnections of the lattice. A main advantage of this is that multiple structures within an atlas can be deformed simultaneously. Forces acting on the separate regions influence the deformations of the other regions. The forces acting on neighboring regions either compete and balance each other out or combine and reinforce each other. These template models typically consist of an external force to drive the template into alignment with an image and a regularization term based on physical principles of elasticity or fluidlike behavior. In these approaches, there is no attempt to infer that portions of the imaged object exhibit elastic or fluidlike behavior or to model actual developmental growth of particular structures; rather, these physical models act as convenient analogies through which to derive a regularization force.

Although template deformation presents an advantage in working with multiple structures, there are some limitations with implementations described in the literature. The method presented by [24] bases the deformation on physical elastic constraints. This model penalizes large nonlinear deformations in the template. In the case of brain image

segmentation as shown in Figure 1.2, different brains may have large nonlinear differences in anatomical structure. To cope with this, Christensen et al. [25] developed a fluid-based model in which the deformation is based on a non-mass-conserving fluid continuum mechanics analogy. Both [24] and [25], however, rely on knowledge of the image contrast due to the use of a squared-error minimization or intensity correlation between the input image and the template. Thus, either a separate template for each possible contrast or a preprocessing step to estimate contrast is required. Also, squared-error minimization and intensity correlation techniques are based on the image intensity gradient, so they do not take full advantage of the region information within a structure.

Another disadvantage with existing template deformation methods is their limited use of prior shape information. The elastic and fluid regularization terms have been shown to be equated to prior probabilities through a Gibbs distribution [1] and can be thought of as capturing the Markov relationship between template elements [27]. In practice, shape variations exhibit non-Markovian properties, but existing template deformation methods are not effective in capturing this type of dependency.

2.4 Information-Based Template Matching

Information-based metrics have been used for alignment since at least 1982, when Mars et al. [28] used the mutual information to estimate time delays in electroencephalogram (EEG) signals. Information-based metrics have been introduced to measure rigid alignment of images by Viola and Wells [29], Wells et al. [30], Collignon et al. [31], Maes et al. [32], and Studholme et al. [33]. Measures proposed include joint entropy, mutual information, mutual information minus joint entropy, and mutual information divided by joint entropy.

More recently, mutual information has been investigated in a nonrigid matching setting. The extension from a rigid-body alignment to non-rigid-body deformation is not straightforward. As mentioned in Section 1.2, the degrees of freedom are much higher for non-rigid-body deformations. This is especially important when using MI, since MI

assumes very little prior knowledge about the contrast and intensities in the image. This increases the possibility of converging to a local extremum.

Hata et al. [34] use mutual information between two images to drive small elastic deformations similar to [24]. Optimization is performed using the stochastic gradient method. The gradient calculation is quadratic in the number of samples used to approximate probabilities; thus, only a subset of the image pixels is used in the calculation. Although large, nonlinear deformations are not allowed, preliminary studies for small, linear deformations are encouraging.

Maintz et al. [35] propose a deformation which first performs a rigid-body registration and then refines the result by performing registration on localized windows. The correspondence measure is based on a probability maximization which can be shown to be similar to maximization of mutual information. Probabilities estimated after rigid registration of the image are assumed to be adequate approximations of the probabilities after local deformation and are not updated. Maintz et al. found a trade-off in the window size; a large window size leads to inability to capture local variations, while a small window size leads to violations of spatial connectivity of the image. Results demonstrate the over-deformation problem described in Section 1.2.

Work by Gaens et al. [36] is similar to that of [35] where local transformations are found by dividing images into regions and translating these regions to increase the local similarity criterion. Gaens, however, updates the intensity probabilities as local deformation occurs. Additionally, Gaussian smoothing is performed to promote spatial connectivity of the image.

2.5 MDL-Based Image Partitioning

Methods based on the minimum description length (MDL) principle [37, 38] have been proposed for image understanding. The MDL principle is a general approach in scientific research in which possible explanations are weighted by their complexity. The first MDL-based image partitioning is credited to Leclerc [37]. Although these methods

are not usually considered deformable methods, they are related to the research presented here because they contain an entropy-related cost. The MDL principle asserts that we choose the simplest explanation consistent with the observed data and prior knowledge about the system. The simplicity or complexity of the system can be measured using entropy or Kolmogorov complexity (see Sections 3.3.1 and 3.3.3).

The typical MDL-based image segmentation procedure assumes a descriptive language with which to describe images, and then chooses the shortest description in terms of that language. The descriptive language used in [37] was a low-order polynomial description of intensity variation within each region and a chain-code-like description of the region boundaries. Other work (see, e.g., [38]) has been done to modify the descriptive language or modify the search scheme for the shortest description. A fundamental difference between MDL-based methods and the information-based methods described in Section 2.4 is that MDL methods measure the entropy with respect to a particular pre-selected descriptive language, while the information-based methods attempt to measure the information or correspondence between two images.

2.6 Deformations Using Shape Information

The deformable methods presented so far rely on local information from neighboring image points in the deformation process. Other deformable methods are based on more global shape information. Particularly relevant to this thesis are methods introduced by Cootes et al. [39, 40], Szekely et al. [41], and Duta and Sonka [42] in which prior shape models are trained using eigenshape analysis and are used to constrain a snakelike deformation. The prior shape models are linear models and shape variations are expected to be within an ellipsoidal region around the average shape in the shape space. As explained in [40], in certain instances this is not the case. Dividing a complex shape into several, more well-behaved, subshapes can be problematic because correlation between points in different subshapes is ignored. Also, topology between subshapes is not necessarily

conserved. A conflict could occur, for example, if image points belong to more than one subshape [42].

Another parametric deformable method is that of deformable superquadrics [43, 44]. Superquadrics are a family of parametric shapes that are symmetric in nature. The deformable superquadric model can deform both locally and globally by incorporating the global shape parameters of a superellipsoid with the local degrees of freedom of a membrane spline in a Lagrangian dynamics formulation [2]. Bardinet et al. [45] combined a parametric model with a volumetric deformation by fitting a deformable superquadric to segmented 3-D cardiac images and then refining the fit using a volumetric deformation technique known as free form deformation (FFD). An FFD can be visualized as a rubber-like box in which the object to be deformed is embedded. The outer sides of the box itself are deformed and the deformations of the box are transmitted to the embedded objects. Because the FFD is volumetric, it allows more than one surface model to be deformed simultaneously. Bardinet requires a known segmentation as an input and then attempts to compactly describe the segmentation.

Christensen et al. [46] and Joshi [47] also proposed methods to capture variations in deformations from a template image. Like the work of Bardinet et al., the shape modeling is descriptive rather than generative; once a template has been deformed using either a semiautomated landmark method or elastic method without shape priors, the resulting transform is then tested against a hypothesis to determine if the anatomy under study is normal or abnormal. Statistics of the template deformation are modeled using Gaussian random fields with independent elements.

CHAPTER 3

RELEVANT TOPICS FROM PROBABILITY AND INFORMATION THEORY

The purpose of this chapter is to introduce some theorems, concepts, and notations related to probability and information theory that will be used later in the thesis. In-depth treatment of these topics can be found in [48, 49, 50, 51, 52, 53] from which this section was derived. Standard proofs are not provided unless they have specific significance to the research presented in the chapter.

3.1 Random Variables

A *random variable* (r.v.) is a mathematical model for a phenomenon that behaves in an unpredictable manner from the viewpoint of the observer. For example, the phenomenon may be a voltage measurement across one component in a circuit, the outcome of a coin toss, the value of a stock market index, or a pixel intensity in an image. The outcome of the event or experiment may be unpredictable for several reasons including computational, modeling, or measurement limitations. Such uncertainty can be conveniently modeled by letting the outcome of such an experiment be an r.v.

Random variables can be either discrete or continuous. Loosely speaking, when the number of possible outcomes is countable, the random variable is considered discrete, and if the range of X is continuous, the r.v. is considered continuous. As a discrete example, let X represent the roll of a die. By convention, upper case letters are used to

represent r.v.'s. The set of all possible outcomes of the r.v. is called the *sample space*, and in this case the sample space is $\Omega_X = \{1, 2, 3, 4, 5, 6\}$. Lower case letters, like x , are used to represent elements of the sample space. A probability measure is defined for all possible subsets of the sample space. The notation used here is $P_X(x) = Pr(X = x)$, which in the die example, is the probability that the outcome of the roll of the die turns up x . For a fair die, $P_X(x) = \frac{1}{6}$, $\forall x \in \Omega_X$ since the outcomes are equally likely. In some cases the notation is further reduced to $P(x)$ with the hope that it is clear from the context which random variable is being referred to.

3.1.1 Probability density functions and probability mass functions

The collection of probability measures on all elements of the sample space is called the *probability mass function* (PMF) for discrete r.v.'s. A similar function defined for continuous r.v.'s is the *probability density function* (PDF). For both continuous and discrete r.v.'s a *cumulative distribution function* (CDF) exists and is defined by

$$F(x) = Pr(X \leq x) \quad (3.1)$$

For discrete r.v.'s it is the case that

$$\sum_{x \in \Omega_X} P(x) = 1 \quad (3.2)$$

and also

$$0 \leq P(x) \leq 1 \quad (3.3)$$

For continuous r.v.'s the probability density function can be defined as

$$f_X(x) = \frac{dF(x)}{dx} \quad (3.4)$$

Although, it is true that

$$\int_{\Omega_X} f(x) = 1 \quad (3.5)$$

it is not the case that

$$0 \leq f(x) \leq 1 \quad (3.6)$$

since although $f(x)$ must be nonnegative for all x , it can be arbitrarily large at a point.

A discrete r.v. that is relevant to this thesis is the *Bernoulli* r.v. A Bernoulli r.v. has two possible outcomes, say 0 or 1. Let $P_X(0) = p$ and $P_X(1) = q$, with $p > 0$, $q > 0$, $p + q = 1$. An example of a Bernoulli r.v. is a coin toss. If the coin is “fair” then $p = q = 0.5$.

Perhaps the most used continuous r.v. model is the Gaussian r.v. The PDF of a Gaussian r.v is described by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (3.7)$$

where μ and σ are called the mean and standard deviation of the random variable. The distribution described by (3.7) is named the *normal* distribution and is abbreviated by $N(\mu, \sigma^2)$.

3.1.2 The Gibbs distribution

The Gibbs distribution is a distribution based on principles of statistical mechanics and probabilities of energy states at thermal equilibrium. Consider a physical system which can reside in any of a large number of possible states. Let P_i denote the probability that the system resides in state i . Also, define $\mathcal{E}_i$ as the *energy* of the system when it is in state i . A fundamental result from statistical mechanics states that when the system is in thermal equilibrium, state i occurs with probability

$$P_i = \frac{1}{Z} \exp\left(-\frac{\mathcal{E}_i}{k_B T}\right) \quad (3.8)$$

where T is the absolute temperature in kelvin, k_B is *Boltzmann's constant*, and Z is a normalizing constant given by

$$Z = \sum_i \exp\left(-\frac{\mathcal{E}_i}{k_B T}\right) \quad (3.9)$$

The principle of the Gibbs distribution has been used in signal processing and computer vision areas to relate system energies to probabilities. In a less formal setting than that of (3.8), probability density functions can be written in the form

$$f(x) = \frac{1}{Z} \exp \left(-\mathcal{E}(x) \right) \quad (3.10)$$

where $\mathcal{E}(x)$ is the *Gibbs potential* for random variable X . One example of a Gibbs potential is the Gaussian potential given by

$$\mathcal{E}(x) = \frac{x^2}{2\sigma^2} \quad (3.11)$$

Similarly, consider random vectors $\mathbf{X}$ and $\mathbf{Y}$ related by

$$\mathbf{X}_i = \mathbf{Y}_i + \eta_i \quad (3.12)$$

where $\{\eta_i\}$ is a set of independent samples from a Gaussian process. Then the energy associated with $P(\mathbf{y}|\mathbf{x})$ is

$$\mathcal{E}(\mathbf{y}|\mathbf{x}) = \frac{1}{2\sigma^2} \sum_i [x_i - y_i]^2 \quad (3.13)$$

3.1.3 Functions of random variables

Random variables can be used in mathematical equations just as nonrandom variables are. The value of an equation that includes a random variable is also random; an r.v. Y can be defined by $Y = g(X)$. If the CDF is known for X , the CDF for Y can be determined by

$$F(y) = Pr(Y \leq y) \quad (3.14)$$

$$= Pr(g(X) \leq y) \quad (3.15)$$

$$= Pr(X \leq g^{-1}(y)) \quad (3.16)$$

assuming g^{-1} exists. Probability mass functions can be defined by

$$P(y) = \sum_{x \in \Omega_X} P(x) \delta(y, g(x)) \quad (3.17)$$

where $\delta(\cdot, \cdot)$ is the Kronecker delta function with property

$$\delta(a, b) = \begin{cases} 1 & a = b, \\ 0 & \text{otherwise} \end{cases} \quad (3.18)$$

For continuous r.v.'s it is possible in many cases to determine $f(y)$ by

$$f_Y(y) = \frac{f_X(x)}{\left| \frac{dg(x)}{dx} \right|} \quad (3.19)$$

Random variables model single events which have a single outcome. Phenomena that have more than one outcome may be modeled with *random vectors* or *random processes*. Random vectors and random processes are indexed collections of random variables. Random processes differ from random vectors in that they have an infinite index which almost always corresponds to time. Random vectors are denoted in boldface type, such as **X**.

3.1.4 Joint and conditional probabilities

When dealing with more than one r.v., joint and conditional probabilities exist. The *joint* probability of X and Y is the probability of two events both occurring, written by

$$P_{XY}(x, y) = Pr(X = x \text{ and } Y = y) \quad (3.20)$$

Two random variables are *independent* when

$$P_{XY}(x, y) = P_X(x)P_Y(y) \quad (3.21)$$

If (3.21) does not hold, then the two random variables are dependent; that is, knowledge of the outcome of one random variable will influence the probability of the other occurring. The probability of X given knowledge of the outcome of Y is a *conditional* probability and is denoted by

$$P_{X|Y}(x|y) = Pr(X = x \text{ given } Y = y) \quad (3.22)$$

In all cases

$$P_{XY}(x, y) = P_{X|Y}(x|y)P_Y(y) \quad (3.23)$$

Probabilities $P_X(x)$ and $P_Y(y)$ themselves are called *marginal* probabilities and can be derived from the joint probability using

$$P_X(x) = \sum_{y \in \Omega_Y} P_{XY}(x, y) \quad (3.24)$$

$$P_Y(y) = \sum_{x \in \Omega_X} P_{XY}(x, y) \quad (3.25)$$

3.1.5 Simple statistics of random variables

A *statistic* is a deterministic value computed about a random variable which lends to an understanding of its uncertain behavior. An often used statistic is the *expected value* or *mean* of an r.v. The expected value $E_X(X)$ of a discrete r.v. is defined as

$$E_X(X) = \sum_{x_i \in \Omega_X} x_i P_X(x_i) \quad (3.26)$$

Other notations used for $E_X(X)$ are μ_X or $E(X)$.

Intuitively $E(X)$ is the average value that samples drawn from P_X will take on. In fact, it can be proven that as the number of samples grows, the average value of the samples will converge to the mean of X . More formally, consider an experiment for which outcomes follow a probability law $P_X(x)$. Let $X_i, i = 1, \dots, n$ denote random variables associated with n outcomes of the experiment, then

$$E(X) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n x_i \quad (3.27)$$

Equation (3.27) is a statement of the *law of large numbers* (see [52], p. 357).

For a limited number of samples, the mean of the r.v. can be estimated by

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n x_i \quad (3.28)$$

Another statistic is the *variance* of a random variable. It is defined mathematically by

$$\text{VAR}(X) = E_X[(X - E(X))^2] = E_X(X^2) - E^2(X) \quad (3.29)$$

The *standard deviation* of a random variable is another popular statistic. It is simply the square root of the variance.

Much of the utility of an r.v. is dependent on knowledge of an associated PDF or PMF. However, in many cases the probability distributions are unknown and must be estimated from measured data. One popular estimation technique is to assume the r.v. follows a probability law that is a function of a few statistical parameters such as the mean and variance. These statistics are then estimated from experimental samples. For example, the noise in image data is often assumed to conform to a normal distribution. The distribution is then characterized by its mean and variance.

3.1.6 Central limit theorem

One justification for the use of a normal probability distribution function in modeling random variables is the central limit theorem. Basically, the central limit theorem says that the distribution of the normalized sum of a large number of mutually independent random variables with zero means and finite variances tends to the normal probability distribution provided that the individual variances are small enough. The following is a more formal statement of this important theorem.

Theorem 3.1 (Central Limit Theorem, [52], p. 213) *Let $X_1, \dots, X_n$ be n mutually independent (scalar) r.v.'s with CDFs $F_1(x_1), F_2(x_2), \dots, F_n(x_n)$, respectively, such that*

$$E[X_k] = 0, \quad \text{Var}[X_k] = \sigma_k^2$$

and let

$$s_n^2 = \sigma_1^2 + \dots + \sigma_n^2$$

If for a given $\epsilon > 0$ and n sufficiently large, the σ_k satisfy

$$\sigma_k < \epsilon s_n^2 \quad k = 1, \dots, n$$

then the normalized sum

$$Y = (X_1 + \dots + X_n)/s_n^2$$

converges to the standard normal PDF, that is, $\lim_{n \rightarrow \infty} f_Y(y) = N(0, 1)$.

3.2 Parzen Window Density Estimation

A popular, nonparametric approach to probability density estimation is the Parzen window technique. This technique is much more flexible than parametric methods since Parzen estimation requires only that the density being estimated be smooth. The general form of the estimate is

$$\hat{f}_X(x) = \frac{1}{N} \sum_{i=1}^N R(x - x_i) \quad (3.30)$$

where $\hat{f}_X(x)$ is an estimate of the probability density at point x , the x_i are the results of N samples drawn independently from the distribution being estimated, and $R(x)$ is a windowing or smoothing function. Equivalently, (3.30) can be rewritten as

$$\hat{f}_X(x) = \frac{1}{Nb_N} \sum_{i=1}^N R_0\left(\frac{x - x_i}{b_N}\right) \quad (3.31)$$

to explicitly show the dependence of $\hat{f}_X(x)$ on the width b_N of the windowing function. The estimate is assured to be a legitimate density function if

$$R(x) \geq 0, \quad \forall x$$

and

$$\int R(x) dx = 1$$

The most common windowing functions are symmetric about the origin, are unimodal, and fall off quickly to zero. Very common windowing functions are the boxcar (rectangular) function and the Gaussian function. The Parzen density estimator essentially computes a local average of the samples. Notice that if R is symmetric about the origin, the estimate at a point can be calculated as a weighted sum over all sample points, where the weighting is determined by the window function centered at the query point.

The technique relies on the fact that the probability p_j that a vector will fall in a region $\mathbb{X}_j$ is given by

$$p_j = \int_{\mathbb{X}_j} f(x) dx \quad (3.32)$$

Thus the p_j represent a smoothed and discretized version of $f(x)$. If we assume that $f(x)$ is continuous and that $\mathbb{X}_j$ is small enough that $f(x)$ does not vary appreciably within it, then

$$p_j \approx f(x_j)V \quad (3.33)$$

where x_j is a point in $\mathbb{X}_j$ and V is the volume of $\mathbb{X}_j$. Thus, the only theoretical restriction on $f(x)$ is that it is smooth.

The Parzen estimate can also be thought of as an estimate of an expectation:

$$\hat{f}(x_0) = \frac{1}{N} \sum_{i=1}^N R(x_0 - x_i) \approx E[R(x_0 - X)] \quad (3.34)$$

As N becomes large we have

$$\lim_{N \rightarrow \infty} \hat{f}(x_0) = E[R(x_0 - X)] \quad (3.35)$$

$$= \int_{-\infty}^{\infty} R(x_0 - x) f(x) dx \quad (3.36)$$

$$= R(x_0) * f(x_0) \quad (3.37)$$

In the limit $\hat{f}(x)$ converges to a convolution between R and the true $f(x)$, so $\hat{f}(x)$ converges to $f(x)$ if and only if $f(x) = R(x) * f(x)$. When this equality holds, $\hat{f}(x)$ becomes an unbiased estimator. For an infinite number of samples this can be accomplished by defining $R(x)$ such that it behaves like a Dirac delta function in the limit.

For a finite number of samples, letting $R(x)$ be a delta function will usually give a useless density estimate. However, under some conditions on $f(x)$, the equality $f(x) = R(x) * f(x)$ still holds for $R(x)$ other than a delta function. When $f(x)$ is frequency-band-limited and $R(x)$ is a low-pass filter such that $f(x)$ can pass undistorted through $R(x)$, the Parzen estimate will be unbiased. Although such a combination is rarely the strict case, it is approximately true in many circumstances. For example, many imaging processes include an additive noise term. The samples used in the estimation can be expressed in terms of another random variable plus uncorrelated noise:

$$X_i = Y_i + \eta \quad (3.38)$$

The density of X_i is equal to the density of Y_i convolved with the density of the noise. For many noise models including Gaussian and chi-squared, this has the effect of smoothing the distribution of Y such that X is a low-pass version of Y 's distribution. In this case, a good Parzen window is one with the same bandwidth as the noise density function.

Because the estimate must be calculated from a finite number of samples, it is instructive to know how the estimate varies with the number of samples. The variance of $\hat{f}(x)$ is

$$\text{VAR}(\hat{f}(x)) = E \left[\left(\hat{f}(x) - E\hat{f}(x) \right)^2 \right] \quad (3.39)$$

$$= E \left[\left(\frac{1}{N} \sum_i R(x - X_i) - E \frac{1}{N} \sum_i R(x - X_i) \right)^2 \right] \quad (3.40)$$

$$= \frac{1}{N^2} E \left[\left(\sum_i R(x - X_i) - E \sum_i R(x - X_i) \right)^2 \right] \quad (3.41)$$

$$= \frac{1}{N} (E R^2(x - X_i) - E^2 R(x - X_i)) \quad (3.42)$$

$$= \frac{1}{N} \int R^2(x - x_i) f(x_i) dx_i - \frac{1}{N} \int R(x - x_i) f(x_i) dx_i \quad (3.43)$$

Equations (3.39) and (3.40) come from (3.29) and (3.30). Equation (3.42) follows from the fact that the variance of the sum of independent variables is the sum of the variances. The result in (3.43) shows that the variance of the estimate varies inversely with the number

of samples, and will converge to zero if $\int R^2(x - x_i)f(x_i)dx_i$ and $\int R(x - x_i)f(x_i)dx_i$ are bounded.

Perhaps a more useful measure of the utility of estimate $\hat{f}(x)$ is the mean squared error from the actual $f(x)$ rather than from its own mean. This can be written as [54]

$$E | \hat{f}(x) - f(x) |^2 = \text{VAR}(\hat{f}(x)) + | E\hat{f}(x) - f(x) |^2 \quad (3.44)$$

So for unbiased estimates of $f(x)$ the mean squared error is equal to the variance of the estimator itself.

It can further be shown [54, 55] that the estimate $\hat{f}(x)$ at a particular point x asymptotically approaches a normal distribution with mean $f(x)$ and variance $f(x)/(Nb_N)$ for many typical windowing functions including boxcar, Gaussian, and triangle. The proof is shown below for the 1-D case using a boxcar function.

Let

$$R_0(\varrho) = \begin{cases} 1 & \text{if } |\varrho| \leq 0.5, \\ 0 & \text{otherwise} \end{cases} \quad (3.45)$$

The probability that the i th sample falls inside the windowing function centered at x , for example, follows the Bernoulli probability law. Let $Y_{x,i}$ define these Bernoulli random variables; then,

$$\hat{f}(x) = \frac{1}{Nb_N} \sum_{i=1}^N Y_{x,i} \quad (3.46)$$

The probability of success of the Bernoulli trial (that $x - b_N/2 < x_i \leq x + B_N/2$) is defined by

$$Pr(Y_{x,i} = 1) = F_X(x + b_N/2) - F_X(x - b_N/2) \approx f(x)b_N$$

where F_X is the CDF of X and $f(x)$ remains approximately constant over the width of the window. If N , x , and b_N are selected such that $Nf(x)b_N - Nf^2(x)b_N^2 > 0$, then another random variable Z can be defined by

$$Z = \frac{\sum (Y_{x,i} - f(x)b_N)}{\sqrt{Nf(x)b_N - Nf^2(x)b_N^2}} \quad (3.47)$$

By the central limit theorem, the distribution of Z approaches $N(0, 1)$ as $N \rightarrow \infty$. Then the distribution of $\hat{f}(x)$ is also normal with mean $f(x)$ and variance $f(x)/(Nb_N)$ if $b_N \rightarrow 0$ and $Nb_N \rightarrow \infty$ as $N \rightarrow \infty$.

For a general windowing function, the estimate is asymptotically normal with mean and variance given by

$$E\hat{f}(x) \approx f(x) \quad (3.48)$$

and

$$\text{VAR}(\hat{f}(x)) \approx \frac{f(x)}{Nb_N} \int R_0^2(\nu) d\nu \quad (3.49)$$

For 2-D estimations,

$$E\hat{f}(x, y) \approx f(x, y) \quad (3.50)$$

and

$$\text{VAR}(\hat{f}(x, y)) \approx \frac{f(x, y)}{Nb_N^2} \int R_0^2(\nu_1, \nu_2) d\nu_1 d\nu_2 \quad (3.51)$$

When considering the shape of the Parzen window, one should refer to previous work done by Epanechnikov [56] and Rosenblatt [54] in which window shapes were studied. The conclusion drawn from [54] is that although an optimal window can be derived, which is expected to reduce the variance most rapidly with increasing sample size, many nonoptimal windows achieve approximately the same rate of decrease.

To illustrate the insensitivity of the estimate to window shape, samples were drawn from a mixture of three Gaussians, followed by the addition of Gaussian noise with standard deviation of 3. Figure 3.1 shows the PDF from which the samples were drawn. The values of the samples were then quantized into 500 bins. The density was estimated from these samples using various Parzen windows. Following the notation of (3.31), let the width of the Parzen window be defined by b_N where

$$R(x) = \frac{1}{b_N} R_0\left(\frac{x}{b_N}\right) \quad (3.52)$$

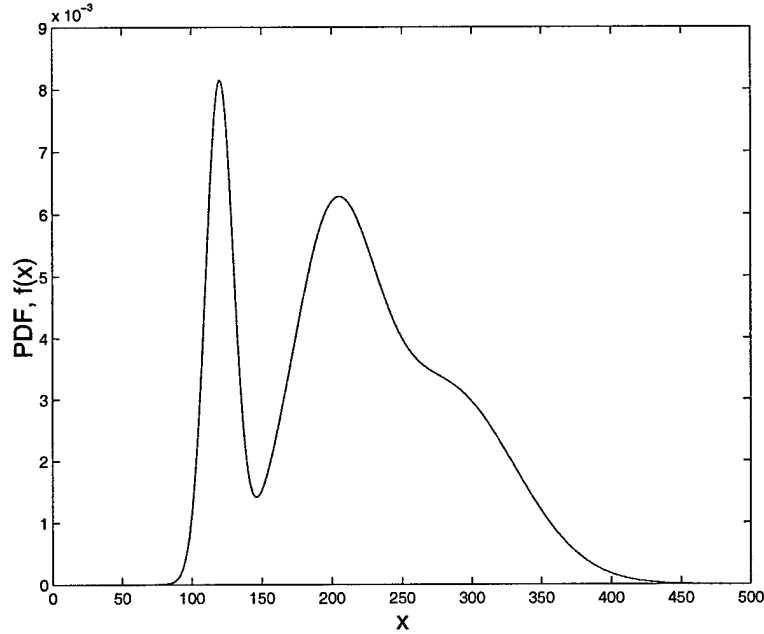


Figure 3.1 Plot of the PDF from which samples were drawn for tests corresponding to Figures 3.2, 3.3, and 3.4.

Figure 3.2 shows the squared error between the estimated and the true distributions as b_N varied. The Gaussian and rectangular (boxcar) windows worked best and had approximately the same minimum. The triangle window was only slightly less effective. It is possible that the Gaussian shape worked best due to the Gaussian nature of underlying distribution and noise.

Figures 3.3 and 3.4 illustrate the change in estimation accuracy as a function of the number of samples N . Figure 3.3 shows the estimation error for a Gaussian Parzen window as a function of the width of the Gaussian pulse when 500 samples were drawn from the distribution shown in Figure 3.1. The number of samples was equivalent to the quantization levels of the samples. Figure 3.4 shows the equivalent experiment when 5000 samples are used. By contrasting Figures 3.3 and 3.4 it is evident that as the number of samples becomes larger, the optimal window width becomes smaller. Less smoothing is necessary since the variance of the estimate at a specific x will become smaller.

The preceding experiments show that for a relatively large number of samples compared, the probability estimate is quite insensitive to both window width and window

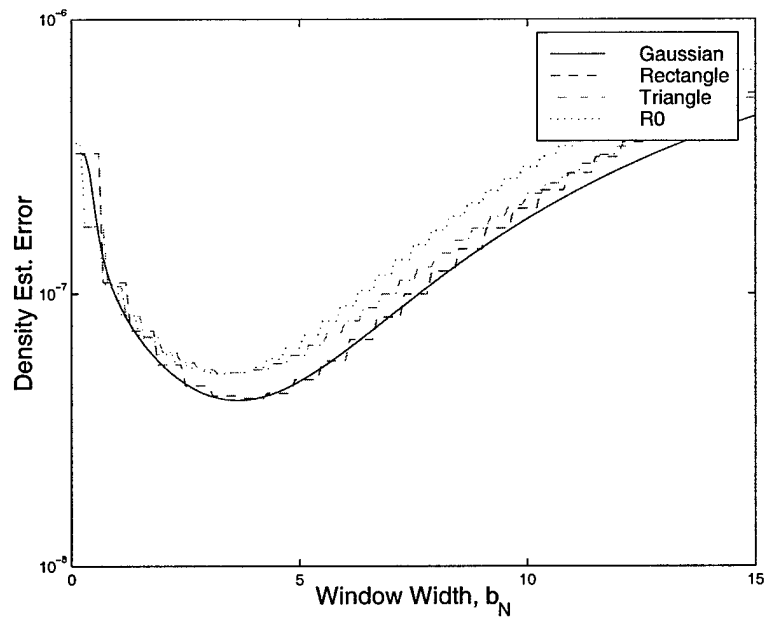


Figure 3.2 Plot of the error between the estimated and actual PDF as a function of Parzen window width b_N for different window shapes. Window shapes shown are Gaussian, rectangle, triangle, and the “optimal” window given in [54].

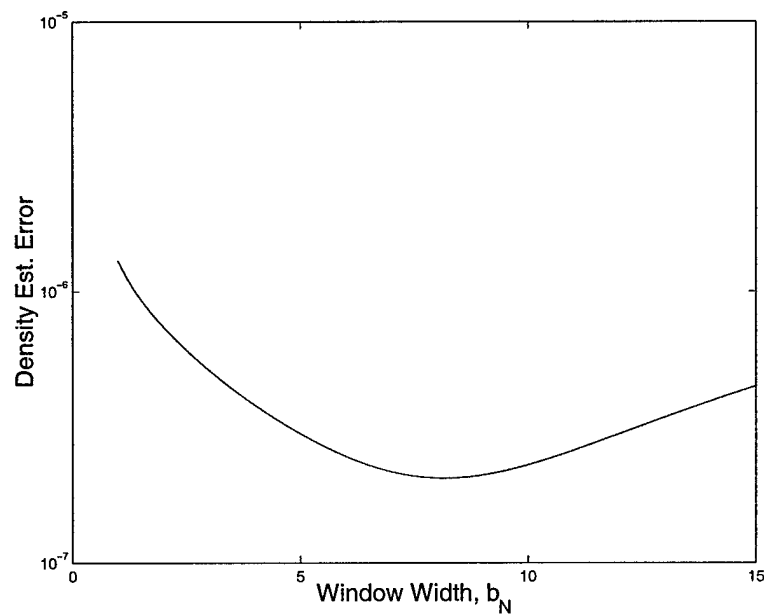


Figure 3.3 Error between the estimated and actual PDFs as a function of window width for a Gaussian window and 500 samples.

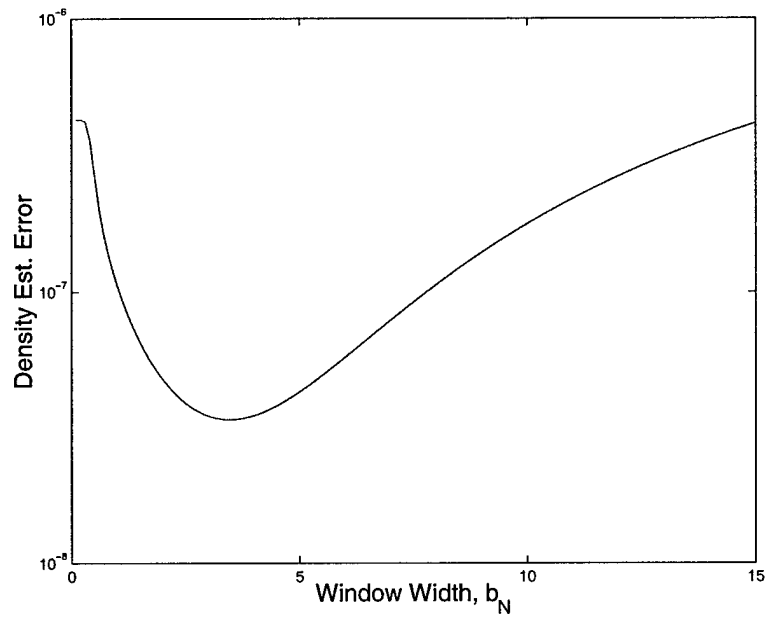


Figure 3.4 Error between the estimated and actual PDFs as a function of window width for a Gaussian window and 5000 samples.

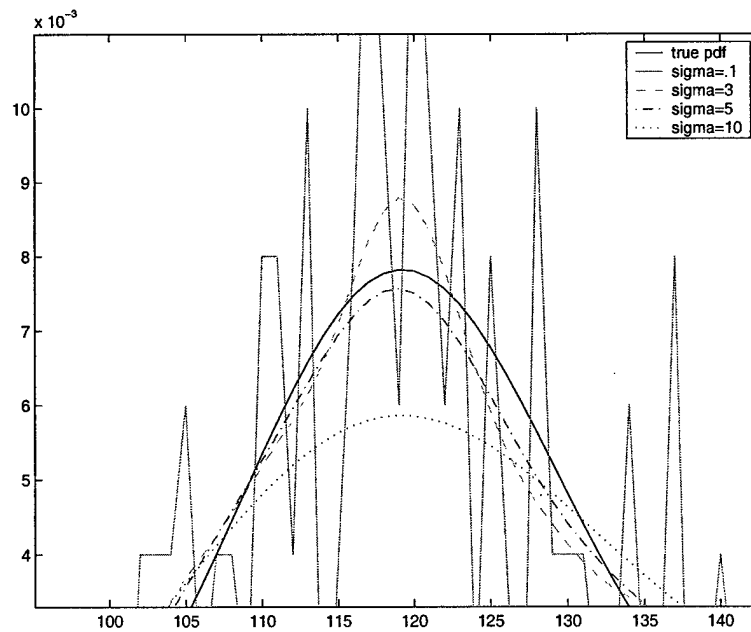


Figure 3.5 Plots of the actual and estimated distributions using a Gaussian Parzen window with several widths.

shape. Throughout this thesis a Gaussian windowing function with a standard deviation of 5 is used.

3.3 Information

The foundations of what is known today as *information theory* were developed by Claude Shannon in 1948 in direct response to the need for electrical engineers to design communication systems that were both efficient and reliable [51]. Information theory, however, has proved useful in far more arenas than communications, such as statistical physics, computer science, statistical inference [48], training of neural networks [51], and image processing [30, 32, 57].

The *information* of an event is the amount of knowledge gained by the occurrence of that event. Consider a discrete r.v. Y with sample space $\Omega_Y \subset \mathcal{R}$. The outcome of Y is y , where y can take any value in Ω_Y with probability $P(y)$. The information gained upon observing the outcome of Y is

$$I(y) = \log \left(\frac{1}{P(y)} \right) \quad (3.53)$$

where $\log$ is usually taken to be base 2. The information gained by the occurrence of a likely event is small since the outcome was in some sense “expected”; however, if an unlikely event occurs, it provides much more information. Thus $I(y)$ measures the “surprise” involved in the occurrence of y .

3.3.1 Entropy

Another useful measure is the expected information or expected surprise associated with a random variable. Let

$$\begin{aligned} H(Y) &= E[I(y)] \\ &= - \sum_{y \in \Omega_Y} P(y) \log P(y) \end{aligned} \quad (3.54)$$

The quantity $H(Y)$ is the *entropy* of the random variable Y . Alternate notations are $H(P_Y)$ and $H_Y(P)$, which emphasize the fact that entropy measures are based on the distribution of the probability function rather than the actual outcome of the random variable.

Entropy is also defined for pairs of random variables. The *joint entropy* of two random variables is

$$H(X, Y) = - \sum_{\forall x \in \Omega_X} \sum_{\forall y \in \Omega_Y} P(x, y) \log P(x, y) \quad (3.55)$$

and the conditional entropy is

$$H(X | Y) = - \sum_{\forall x \in \Omega_X} \sum_{\forall y \in \Omega_Y} P(x, y) \log P(x | y) \quad (3.56)$$

The joint and conditional entropies are related by

$$H(X, Y) = H(X) + H(Y | X) \quad (3.57)$$

Applying this relationship repeatedly results in the chain rule for entropy:

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1) \quad (3.58)$$

3.3.2 Mutual information

The *relative entropy* or *Kullback-Leibler distance* between two probability mass functions $P(y)$ and $Q(y)$ is

$$D(P||Q) = \sum_{y \in \mathcal{Y}} P(y) \log \frac{P(y)}{Q(y)} \quad (3.59)$$

The relative entropy $D(P||Q)$ is a measure of the inefficiency of assuming that the distribution is Q when the true distribution is P .

Using relative entropy, a related concept called mutual information can be defined. The *mutual information* of two random variables $I(X; Y)$ is the relative entropy between the joint distribution $P(x, y)$ and the joint distribution assuming independence,

$P(x)P(y)$.

$$\begin{aligned} I(X; Y) &= D(P(x, y) || P(x)P(y)) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \end{aligned} \quad (3.60)$$

3.3.3 Kolmogorov complexity

Entropy is also related to Kolmogorov complexity. Kolmogorov defined the algorithmic (descriptive) complexity of an object to be the length of the shortest binary computer program that describes the object. Assuming the computer already knows the size $l(y)$ of an object y , then the *conditional Kolmogorov complexity* can be defined as the minimum length of program p such that a computer $\mathcal{U}$ executing p and knowing $l(y)$ will produce output y . Mathematically this is given by

$$K(y | l(y)) = \min_{p: \mathcal{U}(p, l(y))=y} l(p) \quad (3.61)$$

Under certain conditions, the expected value of the conditional Kolmogorov complexity of a random sequence converges to the entropy of the sequence ([48], p. 153).

CHAPTER 4

TEMPLATE ALIGNMENT AND BODY FORCE

The main topic of this chapter is the body force which drives the template into alignment with an image. The chapter starts with a look at the use of MI as an image alignment metric based on work by [30, 32, 53, 58], and then extends the study of MI to the more specific case of non-rigid-body registration.

4.1 Alignment Based on Mutual Information

Consider an imaging process which provides some information about the underlying structure $\mathbf{s}(\mathbf{x})$ of an object in the form of an image $\mathbf{d}(\mathbf{x})$. For 2-D images, $\mathbf{x} = (x, y)$ and takes on values in $\mathbb{X}$, which without loss of generality can be defined as the unit square. Each element of $\mathbf{d}$ takes on a value between 0 and $N_d - 1$. For grayscale images N_d is usually 256. Each element of $\mathbf{s}$ takes on a value between 0 and $N_a - 1$ where N_a is the number of different materials/structures in the scene. The imaging process (see Figure 4.1) can be formulated by

$$\mathbf{d}(\mathbf{x}) = C(\mathbf{s}(\mathbf{x}), \phi) + \eta \quad (4.1)$$

where ϕ is a single vector representing exogenous factors such as lighting, temperature, etc. The imaging function C defines the *contrast* of the image.

To take advantage of the discrete nature of the image and template, an alternate notation for $\mathbf{d}$, $\mathbf{s}$, and $\mathbf{a}$ can be used. For an N_x by N_y element image, let $N_p = N_x N_y$.

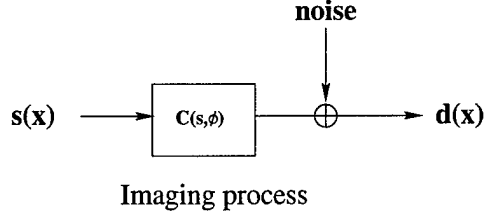


Figure 4.1 A general imaging model. The structure of an underlying scene is converted to an image $\mathbf{d}(\mathbf{x})$ by the imaging process.

Also let i refer to any of N_p specific locations which correspond to an equally spaced grid throughout $\mathbb{X}$. Then $\mathbf{d}$, $\mathbf{s}$, and $\mathbf{a}$ can be considered vectors where, for example, $\mathbf{d} = \{\mathbf{d}(i) : i = 1, 2, \dots, N_p\}$.

The deformable template approach to image understanding is to develop a coordinate displacement field $\mathbf{u} : \mathbb{X} \rightarrow \mathbb{X}$ such that a deformed template $\tilde{\mathbf{a}}(\mathbf{x}) = \mathbf{a}(\mathbf{x} - \mathbf{u}(\mathbf{x}))$ closely approximates $\mathbf{s}$. Due to noise, other uncertainties, and irreversible effects of the imaging process, it may be impossible to know for sure the exact nature of the scene acquired; however, it is assumed that $\mathbf{d}$ contains information about $\mathbf{s}$. The amount of information that can be inferred about $\mathbf{s}$ from $\mathbf{d}$ depends on the nature of C , the amount of uncertainties and noise, as well as prior knowledge about the imaging process and the scene.

For a number of practical problems, C is not known. Within this thesis, alignment measures which do not depend on knowledge of C are called *contrast-tolerant* or *contrast-invariant* metrics. C may not be known for one of several reasons:

- C may be difficult to model.
- The important exogenous factors are unknown or difficult to measure.
- Perhaps it is unknown *a priori* which of many imaging processes was used to generate $\mathbf{d}$.

The method presented herein assumes no specific knowledge about the nature of C , only that the information mutual to both $\tilde{\mathbf{a}}$ and $\mathbf{d}$ is maximal when correctly aligned.

4.1.1 Squared-error and correlation alignment metrics

To better motivate the need for a contrast-tolerant metric, let us first examine other non-contrast-tolerant metrics that have been used to measure image alignment.

For the case that C , $\mathbf{a}$, and ϕ are known, an alignment metric that is optimal in a probabilistic sense is

$$\log P(\mathbf{d} | \mathbf{u}) = \log P(\mathbf{d} | \mathbf{u}, C, \mathbf{a}, \phi) \quad (4.2)$$

$$= \log P(\eta = \mathbf{d} - C(\mathbf{u}(\mathbf{a}), \mathbf{q})) \quad (4.3)$$

The second equality follows from the fact that if C , $\mathbf{a}$, $\mathbf{u}$, and ϕ are all given, then the only randomness of $\mathbf{d}$ is found in η . The function $P(\mathbf{d} | \mathbf{u})$ is called the *likelihood* of $\mathbf{d}$. If η is from an additive white Gaussian noise process and pixels of $\mathbf{d}$ are conditionally independent given C , $\mathbf{a}$, $\mathbf{u}$, and ϕ , then (4.3) reduces to

$$\log P(\mathbf{d} | \mathbf{u}) = -\varrho_1 - \varrho_2 \sum_{i=0}^{N_p} (\mathbf{d}(i) - C(\tilde{\mathbf{a}}(i), \phi))^2 \quad (4.4)$$

Constants ϱ_1 and ϱ_2 are computed from the variance of the noise and are not important to the maximization of the log likelihood. Because $\tilde{\mathbf{a}}$ is a function of $\mathbf{u}$, the optimal displacement field $\mathbf{u}^*$ is found by

$$\mathbf{u}^* = \arg \min_{\tilde{\mathbf{u}}} \sum_{i=0}^{N_p} (\mathbf{d}(i) - C(\tilde{\mathbf{a}}(i), \phi))^2 \quad (4.5)$$

which is a *squared-error minimization* or *least squares* approach.

For large N_p , squared-error minimization is related to a correlation. The summation of (4.5) is proportional to an expectation

$$\sum_{i=0}^{N_p} (\mathbf{d}(i) - C(\tilde{\mathbf{a}}(i), \phi))^2 \propto [D - C(\tilde{A}, \phi)]^2 \quad (4.6)$$

$$= E[D^2] - 2E[DC(\tilde{A}, \phi)] + E[C^2(\tilde{A}, \phi)] \quad (4.7)$$

where r.v.'s D and $\tilde{A}$ are introduced to model uncertainty in the $\mathbf{d}(i)$ and $\tilde{\mathbf{a}}(i)$. The middle term of (4.7) is the correlation between D and $\tilde{A}$. Correlation has been used as an alignment measure in the literature [24].

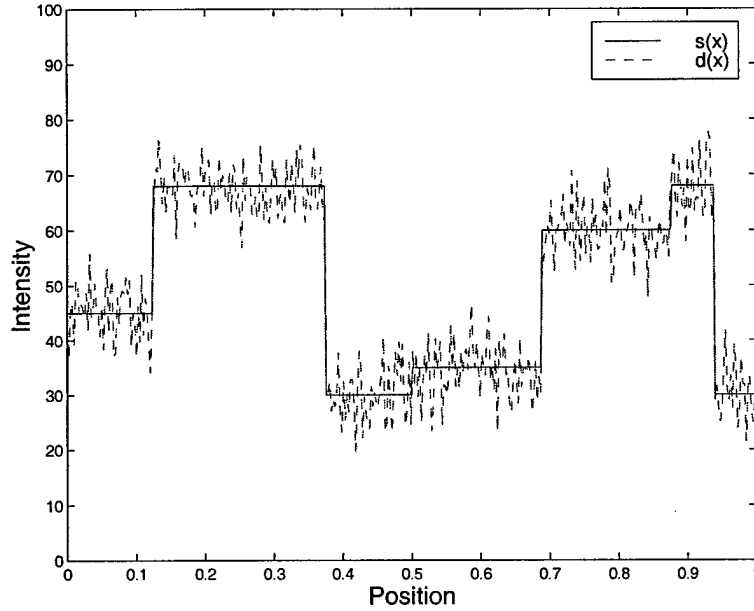


Figure 4.2 A one-dimensional structure and the resulting image based on simply adding noise.

Examples using squared-error alignment

Consider the following simple one-dimensional examples, where $x \in [0, 1]$. Allowable transformations consist of all circular shifts: $u \in [0, 1]$. Let C be the identity function such that

$$d(x) = s(x) + \eta \quad (4.8)$$

where η is white Gaussian noise. Figure 4.2 shows an example s and a possible resulting d based on this C . Figure 4.3 shows the error surface for all possible shifts of s based on (4.5). The minimum squared-error solution is correctly located at zero shift.

A more complicated contrast function would be

$$C(s) = 70 - \frac{(s - 40)^2}{15} \quad (4.9)$$

Figure 4.4 shows s and d for the imaging function of (4.9). Figure 4.5 demonstrates that if C is known, the correct alignment can be found. However, if the wrong C is assumed when finding the alignment, the resulting transform estimate is likely to be far off. For example, assuming d was created from s using the identity imaging function leads to the error curve of Figure 4.6. The minimum error does not lie close to zero shift.

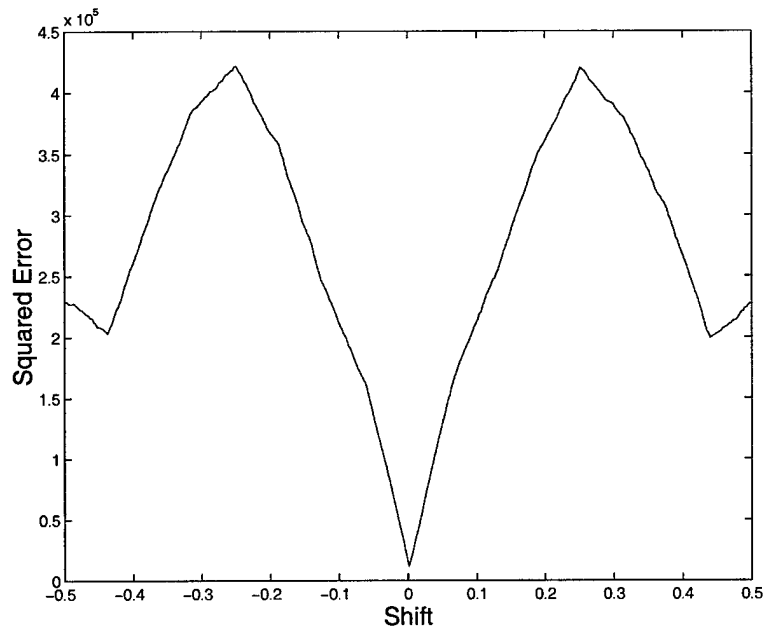


Figure 4.3 Alignment error for all possible shifts of the structure in Figure 4.2. Error is based on the sum of squared intensity differences.

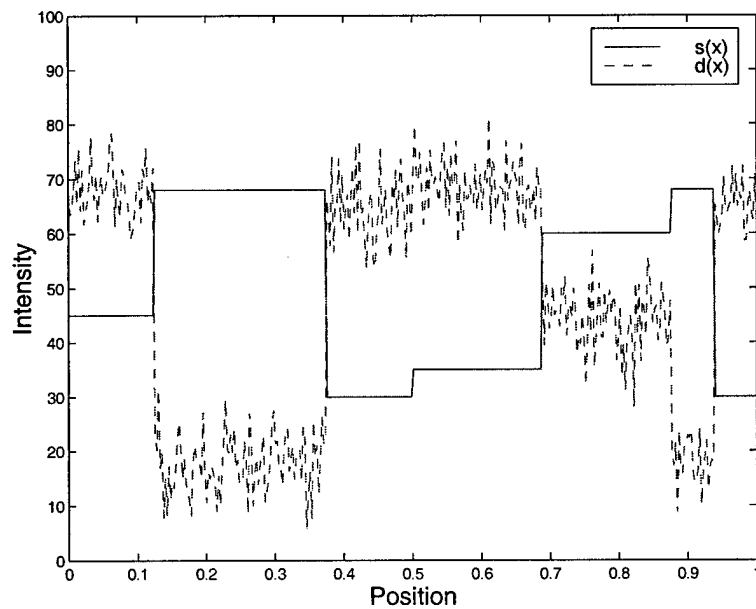


Figure 4.4 A one-dimensional structure and the resulting image based on the imaging function of (4.9).

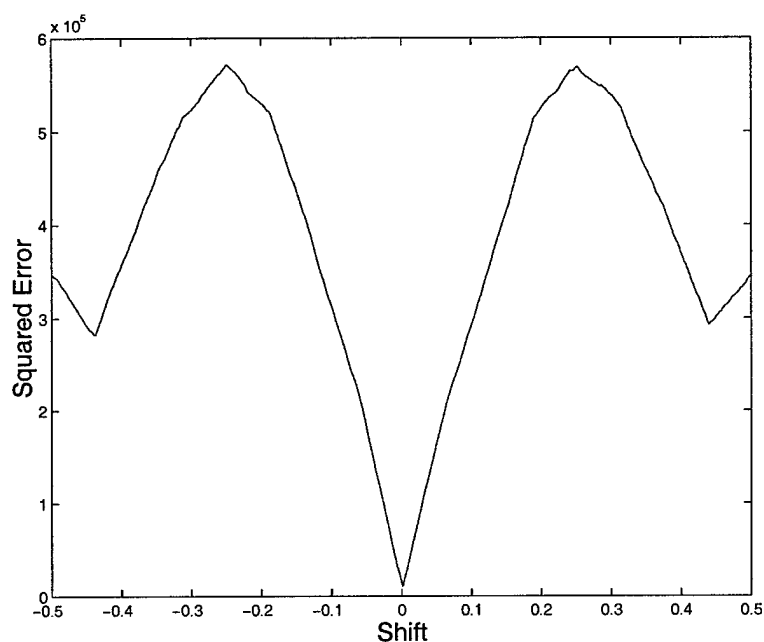


Figure 4.5 Alignment error for all possible shifts of the structure in Figure 4.4. Error is based on the sum of squared intensity differences with the known imaging function of (4.9).

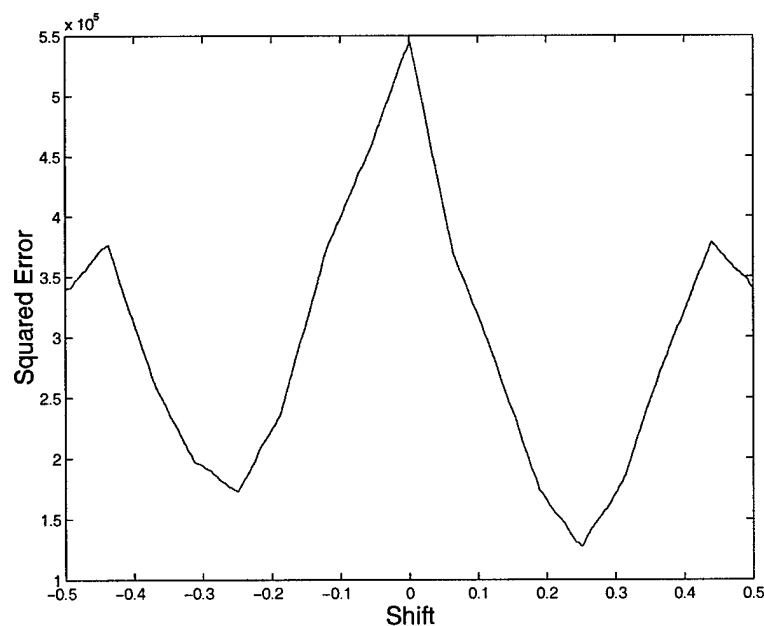


Figure 4.6 Alignment error for the structure in Figure 4.4 when the imaging function was erroneously assumed to be the identity function.

4.1.2 Mutual information for alignment

An alternate measure of image alignment is the estimated mutual information between the image and deformed template. MI can be viewed as a nonparametric approach to alignment because no model is assumed for C . Instead, it is assumed that the MI between the study image and the template is maximal when they are correctly aligned. Implications of this assumption are studied in Section 4.5.1. As mentioned in Section 3.3.1, entropy measures are not based directly on intensity values; rather, the graylevels are treated as symbols and the probabilistic distribution of the symbols determines the entropy and MI. So, MI is a more natural criterion for matching than the squared-error approach when C is not well defined.

Consider two images, $\mathbf{d}$ and $\tilde{\mathbf{a}}$. The mutual information between $\mathbf{d}$ and $\tilde{\mathbf{a}}$ is calculated by introducing random variables D and $\tilde{A}$ with sample spaces $d \in \Omega_D = \{0, \dots, N_d\}$ and $\tilde{a} \in \Omega_{\tilde{A}} = \{0, \dots, N_a\}$, and probabilities empirically determined by

$$P_{D\tilde{A}}(d, \tilde{a}) \approx \frac{1}{N_p} \sum_{m=1}^{N_p} \sum_{n=1}^{N_p} R2(\mathbf{d}(m) - d, \tilde{\mathbf{a}}(n) - \tilde{a}) \quad (4.10)$$

$$P_D(d) \approx \frac{1}{N_p} \sum_{m=1}^{N_p} R1(\mathbf{d}(m) - d) \quad (4.11)$$

$$P_{\tilde{A}}(\tilde{a}) \approx \frac{1}{N_p} \sum_{n=1}^{N_p} R1(\tilde{\mathbf{a}}(n) - \tilde{a}) \quad (4.12)$$

Here $R1$ and $R2$ are appropriate 1D and 2-D Parzen windowing functions. The MI between the images is then

$$I(D; \tilde{A}) = \sum_{\tilde{a}=0}^{N_a-1} \sum_{d=0}^{N_d-1} P_{D\tilde{A}}(d, \tilde{a}) \log \frac{P_{D\tilde{A}}(d, \tilde{a})}{P_D(d)P_{\tilde{A}}(\tilde{a})} \quad (4.13)$$

For this thesis, notation is relaxed slightly so that $I(\mathbf{d}; \tilde{\mathbf{a}}) = I(D, \tilde{A})$ signifies the MI between images $\mathbf{d}$ and $\tilde{\mathbf{a}}$ using (4.10) through (4.13) and implicit r.v.'s D and $\tilde{A}$. Similar notation for the entropy of images is also used to allow $H(\mathbf{d}, \tilde{\mathbf{a}}) = H(D, \tilde{A})$, $H(\mathbf{d}) = H(D)$, and $H(\tilde{\mathbf{a}}) = H(\tilde{A})$.

In theory, the estimated probabilities in (4.10) through (4.12) are continuous PDFs when the windowing functions are continuous functions. However, closed form solutions

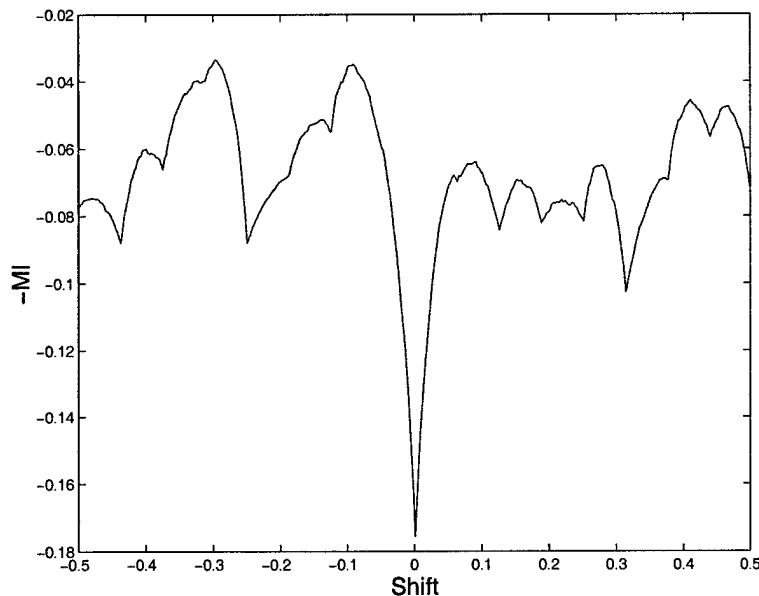


Figure 4.7 Mutual information alignment error curve for the example structure and image shown in in Figure 4.2.

for entropy and MI based on continuous distributions do not exist in general. Although there is a closed form solution for Gaussian densities, mixtures of Gaussians must be calculated empirically. Because of this and the discrete nature of the images, there is no real advantage in maintaining continuous probabilities. Thus, for convenience the probabilities are quantized and expressed in a discrete form.

MI can be used to determine the alignment of the simple 1D examples of Section 4.1.1. Figures 4.7 and 4.8 demonstrate the ability of MI to find the correct alignment for imaging functions illustrated in Figures 4.2 and 4.4, respectively. MI locates the correct alignment in both cases without knowledge of the imaging function.

The gain in generality of the metric is not without cost. By comparing Figures 4.7 and 4.8 with 4.3 and 4.5, it is evident that the region of attraction to the minimum (the width of the main downward spike) is smaller for MI.

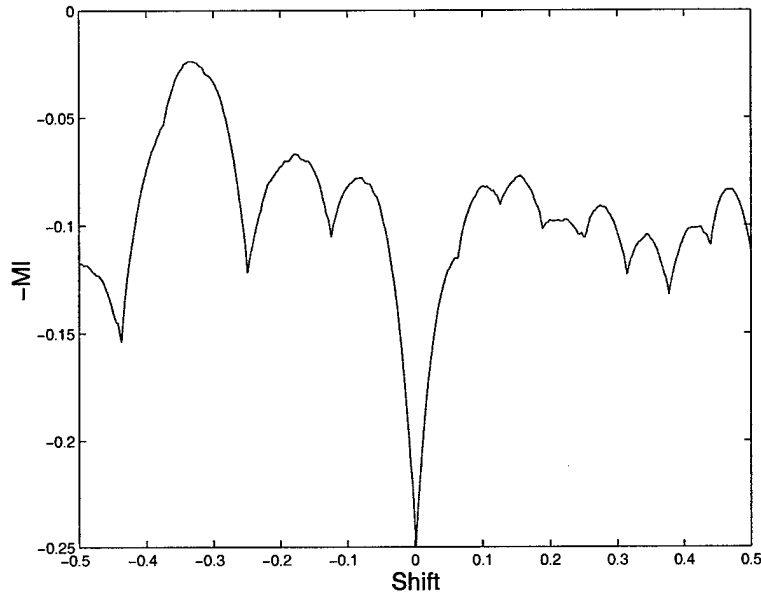


Figure 4.8 Mutual information alignment error curve for the structure and image shown in Figure 4.4.

4.1.3 Parzen windowing for MI

In many implementations, the empirical MI calculation is done using Kronecker delta functions for the Parzen windows used in (4.10) - (4.12). The probabilities are then just the normalized histograms of the images and the joint histogram of the two images. This simplifies analysis, implementation, and understanding of the deformation procedure, but other choices of Parzen windows may lead to better results. The major reason for this is that the MI approach treats the image graylevels as symbols. There is no relative distance between symbols. In symbol space, symbol 1 is not closer to (or farther from) symbol 2 than it is to (from) 100, but in intensity space 1 and 2 are very close, and 1 and 100 are clearly distinguishable.

Particular situations in which a relative distance may be useful are the cases of additive noise and partial volume effects. With additive noise and only a finite number of samples, the probability distributions may not be sufficiently estimated to show the true relationship between a symbol and the same symbol slightly corrupted by white noise.

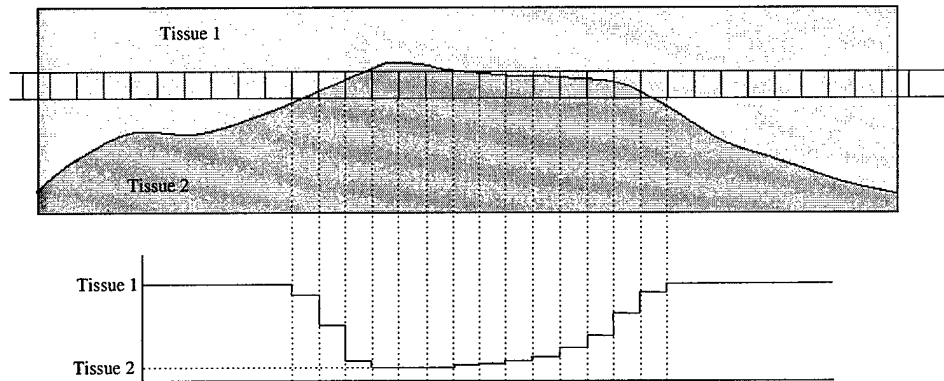


Figure 4.9 Visual explanation of the partial volume effect. Due to finite resolution of the imaging process, an image pixel may contain information from more than one region. (Adapted from [59].)

Partial volume effects are the result of a discrete image element lying on a region boundary. The intensity of the image contains information from two different structures. This is illustrated in Figure 4.9. As a 2-D image slices a 3-D object, pixels overlap more than one region.

Parzen window estimation may reduce noise and partial volume effects. The smoothing property of the Parzen window imposes a relationship between similar graylevels. Figure 4.10 shows a 1-D image with the horizontal axis showing the image position index and the vertical axis showing the intensity (0-255). Assume that the gradual gradient from low to high intensity is a result of partial volume effects. The template is a two valued image which can be shifted along the horizontal axis to divide the study image into two parts. Using MI with normalized histograms leads to the division shown in Figure 4.10(a) with about half of the distinct graylevels on each side of the line. Figure 4.10(b) shows the same experiment but using a Gaussian window to estimate the probabilities. The result is that the image is divided so that all intensities above 126 are on the right side of the line. Based on partial volume effects, the division of Figure 4.10(b) is much more accurate than that of Figure 4.10(a).

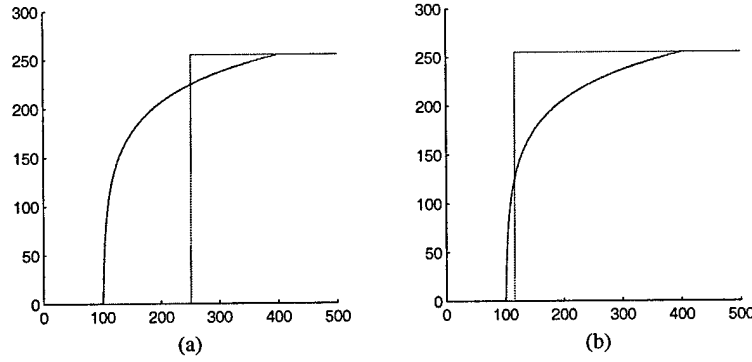


Figure 4.10 1-D image example showing the utility of Parzen window probability estimation in the presence of partial volume effects. The curved line is a 1-D image showing partial volume effects in the transition from intensity 0 to 255. Probability estimation based on histogram normalization leads to the division shown in (a). Parzen window estimation leads to the more accurate division shown in (b).

4.2 Using the MI to Find a Body Force

For the simple 1-D examples above, locating the correct alignment was a simple problem of computing the MI for all possible shifts of the template. For multi-dimensional, non-rigid-body deformations the search space is prohibitively large, precluding an exhaustive search of the transform space. Instead, a gradient search method becomes more practical. The gradient of the MI between the template and study is used to compute a body force $\mathbf{b}_d(\mathbf{x})$ which acts on the template to drive it into alignment with the study image data.

Since $\mathbf{d}$ does not change throughout the deformation process, maximizing MI is equivalent to minimizing a conditional entropy:

$$\max I(\mathbf{d}; \tilde{\mathbf{a}}) = \max \{H(\mathbf{d}) + H(\tilde{\mathbf{a}}) - H(\mathbf{d}, \tilde{\mathbf{a}})\} \quad (4.14)$$

$$= \max \{H(\tilde{\mathbf{a}}) - H(\mathbf{d}, \tilde{\mathbf{a}})\} \quad (4.15)$$

$$= \min H(\mathbf{d} | \tilde{\mathbf{a}}) \quad (4.16)$$

The quantity $H(\mathbf{d}, \tilde{\mathbf{a}})$ is a measure of the uncertainty introduced by the imaging process and can be called the *imaging entropy*.

The body force is computed by taking the localized variation of $H(\mathbf{d} \mid \tilde{\mathbf{a}})$ with respect to $\mathbf{u}$. Recall that $\tilde{\mathbf{a}}$ is a function of $\mathbf{u}$ since $\tilde{\mathbf{a}}(\mathbf{x}, t) = \mathbf{a}(\mathbf{x} - \mathbf{u}(\mathbf{x}), t)$.

$$\mathbf{b}_d(\mathbf{x}, \mathbf{u}(\mathbf{x}, t)) = -\nabla_{\mathbf{u}} H(\mathbf{d} \mid \tilde{\mathbf{a}}) \quad (4.17)$$

$$\begin{aligned} &= \sum_d \sum_{\tilde{a}} \nabla_{\mathbf{u}} [P_{D\tilde{A}}(d, \tilde{a}) \log P_{D\tilde{A}}(d, \tilde{a})] \\ &\quad - \sum_{\tilde{a}} \nabla_{\mathbf{u}} [P_{\tilde{A}}(\tilde{a}) \log P_{\tilde{A}}(\tilde{a})] \end{aligned} \quad (4.18)$$

where $\nabla_{\mathbf{u}}$ is the gradient with respect to the displacement field. Letting $\hat{x}_i : i = 1, 2$ be the orthogonal unit vectors specifying the 2-D Cartesian space of the images, (4.18) is rewritten as

$$\begin{aligned} \mathbf{b}_d(\mathbf{x}, \mathbf{u}(\mathbf{x}, t)) &= \sum_{i=1}^2 \hat{x}_i \left(\sum_{\tilde{a}=0}^{N_a-1} \sum_{d=0}^{N_d-1} \frac{\partial}{\partial u_{\hat{x}_i}(\mathbf{x})} [P_{D\tilde{A}}(d, \tilde{a}) \log P_{D\tilde{A}}(d, \tilde{a})] \right. \\ &\quad \left. - \sum_{\tilde{a}=0}^{N_a-1} \frac{\partial}{\partial u_{\hat{x}_i}(\mathbf{x})} [P_{\tilde{A}}(\tilde{a}) \log P_{\tilde{A}}(\tilde{a})] \right) \\ &= \sum_{i=1}^2 \hat{x}_i \left(\sum_{\tilde{a}=0}^{N_a-1} \sum_{d=0}^{N_d-1} [1 + \log P_{D\tilde{A}}(d, \tilde{a})] \frac{\partial}{\partial u_{\hat{x}_i}(\mathbf{x})} P_{D\tilde{A}}(d, \tilde{a}) \right. \\ &\quad \left. - \sum_{\tilde{a}=0}^{N_a-1} [1 + \log P_{\tilde{A}}(\tilde{a})] \frac{\partial}{\partial u_{\hat{x}_i}(\mathbf{x})} P_{\tilde{A}}(\tilde{a}) \right) \end{aligned} \quad (4.19)$$

The final equality follows from the relationship

$$\frac{\partial}{\partial u_{\hat{x}_i}(\mathbf{x})} (P \log P) = (1 + \log P) \frac{\partial}{\partial u_{\hat{x}_i}(\mathbf{x})} (P) \quad (4.20)$$

Since the probability distributions are discrete rather than continuous, the partial derivatives are approximated with differences based on the four locations immediately surrounding $\mathbf{x}$. Also for the discrete case, $\frac{\partial}{\partial u_{\hat{x}_i}(\mathbf{x})} (P)$ becomes $\frac{1}{N_p}$, and (4.19) is reduced to

$$\begin{aligned} \mathbf{b}_d(\mathbf{x}, \mathbf{u}(\mathbf{x}, t)) &\approx \frac{1}{N_p} \sum_{i=1}^2 \hat{x}_i \log \left(\frac{[P_{D\tilde{A}}(d(\mathbf{x}), \tilde{\mathbf{a}}(\mathbf{x} - \hat{x}_i)) + \frac{1}{2N_p}]}{[P_{D\tilde{A}}(d(\mathbf{x}), \tilde{\mathbf{a}}(\mathbf{x} + \hat{x}_i)) + \frac{1}{2N_p}]} \right. \\ &\quad \left. \frac{[P_{\tilde{A}}(\tilde{\mathbf{a}}(\mathbf{x} + \hat{x}_i)) + \frac{1}{2N_p}]}{[P_{\tilde{A}}(\tilde{\mathbf{a}}(\mathbf{x} - \hat{x}_i)) + \frac{1}{2N_p}]} \right) \end{aligned} \quad (4.21)$$

The factor $\frac{1}{2N_p}$ is also a function of the discrete difference approximation to the continuous derivative. Equation (4.20) requires the value of P at the point the derivative is taken. In this case it is taken to be halfway between the two values used in the difference.

The computational complexity of calculating $\mathbf{b}_d$ for all $\mathbf{x}$ using (4.21) is linear with respect to the image size. Other proposed MI gradient methods have a cost which is quadratic with respect to the number of samples used to calculate the MI gradient [30].

The preceding derivation of (4.21) assumes that an incremental change in $\mathbf{u}$ at $\mathbf{x}$ only changes the probability mass functions for graylevel values $\tilde{a} = \tilde{\mathbf{a}}(\mathbf{x})$, $\tilde{\mathbf{a}}(\mathbf{x} \pm \hat{x}_i)$ and $d = \mathbf{d}(\mathbf{x})$. This assumption is not strictly true when the Parzen window technique is used to estimate the probability mass functions. However, (4.21) is still a good approximation for most practical window functions including the Gaussian function. The dominant terms of (4.21) will be those for which changes in $P(d, \tilde{a})$ and $P(\tilde{a})$ are large with respect to $\mathbf{u}$. Due to the smoothing nature of the windowing function, the contributing value of each of the dominant terms will be approximately equal to the actual value calculated by (4.21). Thus the actual gradient will be approximately related to the value calculated in (4.21) by a constant.

To test the utility of (4.21) with Parzen windowing, joint and marginal probability mass functions were estimated for a template and cryosection image using a Gaussian window. The Gaussian window was applied to smooth only the graylevel values of the input image and not those of the template. The result of (4.21) was then compared with the exact MI gradient calculation. Comparisons were done for Gaussian window widths of $\sigma = 5$ and $\sigma = 10$. The maximum and average magnitude differences relative to the maximum magnitude of the body force are shown in Table 4.1. Values which were identical using (4.21) and the exact calculation were not included in the average calculation.

In order for the gradient method to be feasible, the path leading to the global minimum (maximum) must be smooth so the search will not be attracted to a local extrema. To examine the occurrence of local extrema, the MI between two images was measured as one was scaled smaller and larger. Figure 4.11 shows the two images and the resulting

Table 4.1 Measurements of the utility of (4.21) when Parzen windowing is used. Comparison is made between the MI gradient calculation using (4.21) and the exact calculation. Maximum and average magnitude differences relative to the maximum body force magnitude are shown.

	var=100	var=25
b Max difference	8.5%	7.8%
b Ave difference	1.0%	.5%

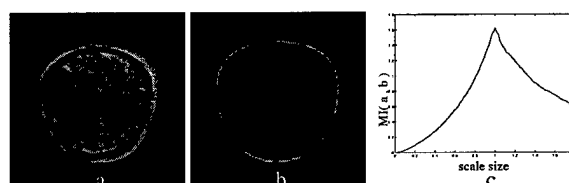


Figure 4.11 Plot of the MI between images (a) and (b) as (b) was scaled smaller and larger.

plot of the MI vs. scale value. As is seen by the graph, the resulting plot is quite smooth.

To examine further the possibility of local extrema, a test was performed similar to one performed by Maes et al. [32] wherein the pixels of the images in Figure 4.11 were randomly mixed so that the marginal intensity distributions and self entropies remained unchanged for the full-scale images, but the images had a less defined structure. Figure 4.12 shows the modified images and the resulting plot of the MI with respect to scaling. The global maximum occurs at the same point, but there are many local maxima and the region of attraction is much smaller. This test demonstrates that a gradient search method will work better on well structured images.



Figure 4.12 Plot of the MI between images (a) and (b) as (b) was scaled smaller and larger. In this case, images do not have a well defined structure, so the plot is not very smooth, and the region of attraction for the global maximum is small.

4.3 MI Bias and Variance

Because an image contains only a finite number of pixels, the Parzen probability estimate will have nonzero variance. One may ask how this affects the MI calculation. It turns out that since entropy is a nonlinear (in fact concave) function of the probability distribution, the entropy estimate is biased. Work on quantifying both the bias and variance of MI and entropy estimates from a finite number of samples has been done by Moddemeijer [60] for histogram-based probability estimates of continuous data. Histogram-based probability estimation is essentially equivalent to using a rectangular Parzen window with unit width. This work is summarized below with slight modification for discrete data, followed by its extension to nonrectangular Parzen windows.

The joint and marginal histograms of two images $\tilde{\mathbf{a}}$ and $\mathbf{d}$ are defined as

$$k_{d,\tilde{a}} = \sum_{m=1}^{N_p} \sum_{n=1}^{N_p} \delta(\mathbf{d}(m) - d, \tilde{\mathbf{a}}(n) - \tilde{a}) \quad (4.22)$$

$$k_{d,\cdot} = \sum_{m=1}^{N_p} \delta(\mathbf{d}(m) - d) \quad (4.23)$$

$$k_{\cdot,\tilde{a}} = \sum_{n=1}^{N_p} \delta(\tilde{\mathbf{a}}(n) - \tilde{a}) \quad (4.24)$$

Let uppercase variables $K_{d,\tilde{a}}$, $K_{d,\cdot}$, and $K_{\cdot,\tilde{a}}$ be random variables which represent the uncertainty of the histogram values for finite N_p . The distribution of $K_{d,\tilde{a}}$ for any fixed $\tilde{a}$ and d follows the binomial probability law:

$$E(K_{d,\tilde{a}}) = N_p P(d, \tilde{a}) \quad (4.25)$$

and

$$\text{COV}(K_{d,\tilde{a}}, K_{\gamma,\delta}) = \begin{cases} N_p P(d, \tilde{a})(1 - P(d, \tilde{a})) & \text{if } \tilde{a} = d \text{ and } \gamma = \delta \\ -N_p P(d, \tilde{a})P(\gamma, \delta) & \text{otherwise} \end{cases} \quad (4.26)$$

where $P(d, \tilde{a})$ is the “true” joint probability of random variables D and $\tilde{A}$. The joint entropy and MI of $\tilde{A}$ and D can be estimated using

$$\hat{H}(D, \tilde{A}) = - \sum_{d, \tilde{a}} \frac{k_{d, \tilde{a}}}{N_p} \log \frac{k_{d, \tilde{a}}}{N_p} \quad (4.27)$$

$$\hat{I}(D; \tilde{A}) = - \sum_{d, \tilde{a}} \frac{k_{d, \tilde{a}}}{N_p} \log \frac{k_{d, \tilde{a}} N_p}{k_{\tilde{a}, \cdot} k_{\cdot, d}} \quad (4.28)$$

4.3.1 Bias

As previously stated, the estimators in (4.27) and (4.28) are biased. In the case of discrete r.v.’s, this is due to the finite sample size and is referred to as N -bias in [60]. The expected N -bias can be calculated exactly since the distribution of $K_{d, \tilde{a}}$ is known. For joint entropy this is

$$N\text{-bias} = H(D, \tilde{A}) - E(\hat{H}(D, \tilde{A})) \quad (4.29)$$

$$= H(D, \tilde{A}) - \sum_{d, \tilde{a}} \sum_{k=0}^{N_p} P(k | P(d, \tilde{a})) \left(\frac{k}{N_p} \log \frac{k}{N_p} \right) \quad (4.30)$$

Substitution of the binomial distribution leads to

$$\begin{aligned} N\text{-bias} = H(D, \tilde{A}) - \sum_{d, \tilde{a}} \sum_{k=0}^{N_p} \binom{N_p}{k} P(d, \tilde{a})^k (1 - P(d, \tilde{a}))^{N_p - k} \\ \times \left(\frac{k_{d, \tilde{a}} N_p}{\log \frac{k_{d, \tilde{a}} N_p}{k_{\tilde{a}, \cdot} k_{\cdot, d}}} \right) \end{aligned} \quad (4.31)$$

The N -bias as calculated by (4.31) is always positive due to the concave nature of $-p \log p$; therefore, the entropy estimate is expected to underestimate the actual.

4.3.2 Variance

Direct calculation of the variance of the entropy estimate using the binomial distribution is much more complicated. Instead, an approximation is obtained using a Taylor

expansion. The Taylor expansion of (4.27) is

$$\begin{aligned} \hat{H}(D, \tilde{A}) = \sum_{d, \tilde{a}} \left\{ -P(d, \tilde{a}) \log P(d, \tilde{a}) - \left(\frac{1}{N_p} + \frac{1}{N_p} \log P(d, \tilde{a}) \right) \right. \\ \left. \times (K_{d, \tilde{a}} - N_p P(d, \tilde{a})) + \mathcal{O}\left(\frac{1}{N_p^2}\right) \right\} \end{aligned} \quad (4.32)$$

and the variance of $\hat{H}(D, \tilde{A})$ is

$$\text{VAR}(\hat{H}(D, \tilde{A})) = E(\hat{H}^2(D, \tilde{A})) - E^2(\hat{H}(D, \tilde{A})) \quad (4.33)$$

Substitution of (4.32) into (4.33) gives

$$\begin{aligned} \text{VAR}(\hat{H}(D, \tilde{A})) = \sum_{d, \tilde{a}} \sum_{\gamma, \delta} \left(\frac{1}{N_p} + \frac{1}{N_p} \log P(d, \tilde{a}) \right) \cdot \left(\frac{1}{N_p} + \frac{1}{N_p} \log P_{\gamma, \delta} \right) \\ \times E(K_{d, \tilde{a}} - N_p P(d, \tilde{a}))(K_{cd} - N_p P_{\gamma, \delta}) \\ - \left(\sum_{d, \tilde{a}} \left(\frac{1}{N_p} + \frac{1}{N_p} \log P(d, \tilde{a}) \right) E(K_{d, \tilde{a}} - N_p P(d, \tilde{a})) \right)^2 + \mathcal{O}\left(\frac{1}{N_p^2}\right) \end{aligned} \quad (4.34)$$

Further substitution of (4.26) gives

$$\begin{aligned} \text{VAR}(\hat{H}(D, \tilde{A})) = \frac{1}{N_p} \left(\sum_{d, \tilde{a}} P(d, \tilde{a}) \log^2 P(d, \tilde{a}) - \left(\sum_{d, \tilde{a}} P(d, \tilde{a}) \log P(d, \tilde{a}) \right)^2 \right) \\ + \mathcal{O}\left(\frac{1}{N_p^2}\right) \end{aligned} \quad (4.35)$$

Similar derivation yields

$$\begin{aligned} \text{VAR}(\hat{I}(D, \tilde{A})) = \frac{1}{N_p} \left(\sum_{d, \tilde{a}} P(d, \tilde{a}) \log^2 \frac{P(d, \tilde{a})}{P(\tilde{a})P(d)} - \left(\sum_{d, \tilde{a}} P(d, \tilde{a}) \log \frac{P(d, \tilde{a})}{P(\tilde{a})P(d)} \right)^2 \right) \\ + \mathcal{O}\left(\frac{1}{N_p^2}\right) \end{aligned} \quad (4.36)$$

Equations (4.35) and (4.36) can be rewritten as

$$\text{VAR}(\hat{H}(D, \tilde{A})) = \frac{1}{N_p} \text{VAR}(\log P(d, \tilde{a})) + \mathcal{O}\left(\frac{1}{N_p^2}\right) \quad (4.37)$$

and

$$\text{VAR}(\hat{I}(D, \tilde{A})) = \frac{1}{N_p} \text{VAR} \left(\log \frac{P(d, \tilde{a})}{P(d)P(\tilde{a})} \right) + \mathcal{O} \left(\frac{1}{N_p^2} \right) \quad (4.38)$$

which clearly show that the variance is, to a first approximation, a function of only the number of samples and the underlying distribution.

4.3.3 Generalization for Parzen windows

The N -bias calculation can be extended to predict the MI and entropy bias for the case that a more general Parzen window is used to estimate the probability distributions rather than the Kronecker delta function in (4.22)-(4.24). The resulting entropy or MI estimate has two bias components rather than the single N -bias component. Although smaller, the N -bias remains. The other is based on the fact that the Parzen window inherently smoothes the underlying distribution. This bias component will be called S -bias.

Recall from Section 3.2 that for the Parzen window technique, the probability estimate for a specific $(d, \tilde{a})$ approaches a normal distribution with mean

$$E \left(\frac{K_{d,\tilde{a}}}{N_p} \right) \approx P(d, \tilde{a}) \quad (4.39)$$

and variance

$$\text{VAR} \left(\frac{K_{d,\tilde{a}}}{N_p} \right) \approx P(d, \tilde{a}) \sum_{\nu_1, \nu_2} R^2(\nu_1, \nu_2) \quad (4.40)$$

Equation (4.40) shows that the variance of $P(k | P(d, \tilde{a}))$ is affected by the energy in the Parzen window. Applying a correction to the variance due to the nonunit energy of the windowing function leads to

$$\begin{aligned} N\text{-bias} \approx H(D, \tilde{A}) - \sum_{\nu_1, \nu_2} R^2(\nu_1, \nu_2) \sum_{d, \tilde{a}} \sum_{k=0}^{N_p} \binom{N_p}{k} P(d, \tilde{a})^k (1 - P(d, \tilde{a}))^{N_p-k} \\ \times \left(\frac{k_{d,\tilde{a}} N_p}{\log k_{\tilde{a},d} k_{d,d}} \right) \end{aligned} \quad (4.41)$$

Although the estimate is still expected to be less than the actual entropy, typical Parzen windows will have energy less than unity and will reduce the N -bias. For example, a 1-D Gaussian-shaped Parzen window with standard deviation 5 will have an energy of 0.056, so the N -bias will be reduced by approximately 94%.

The second bias introduced by the Parzen window, the S -bias, is essentially due to the nonzero Parzen window width required for a finite number of samples. As discussed in Section 3.2, the Parzen density estimate actually returns $P(d, \tilde{a}) * R(d, \tilde{a})$, so the Parzen-based entropy estimate is

$$\hat{H}(P(d, \tilde{a})) = H(P(d, \tilde{a}) * R(d, \tilde{a})) \quad (4.42)$$

and the S -bias is computed by

$$S\text{-bias} = H(P(d, \tilde{a}) * R(d, \tilde{a})) - H(P(d, \tilde{a})) \quad (4.43)$$

Note that the S -bias is not a function of N_p , only of $P(d, \tilde{a})$ and R , and since

$$H(P(d, \tilde{a})) \leq H(P(d, \tilde{a}) * R(d, \tilde{a})) \quad (4.44)$$

the S -bias leads to an overestimation of the entropy or an underestimation of the MI.

Inspection of (4.35) reveals that, to a first-order approximation, the variance is largely only a function of the true underlying probability and the number of samples acquired. Thus, the Parzen window will have a limited effect on the variance of the estimation. Equation (4.35) can be rewritten as

$$\begin{aligned} \text{VAR}(\hat{H}(D, \tilde{A})) = \frac{1}{N_p} & \left(\sum_{d, \tilde{a}} (P(d, \tilde{a}) * R) \log^2(P(d, \tilde{a}) * R) \right. \\ & \left. - \left(\sum_{d, \tilde{a}} (P(d, \tilde{a}) * R) \log(P(d, \tilde{a}) * R) \right)^2 \right) + \mathcal{O}\left(\frac{1}{N_p^2}\right) \end{aligned} \quad (4.45)$$

Thus as the window width increases the variance will decrease.

Figure 4.13 shows actual entropy estimates as a function of N_p , based on samples from a PMF, as well as the expected value of the entropy estimates, based on the N -bias calculation. The N -bias is calculated according to (4.31). The solid, smooth curve

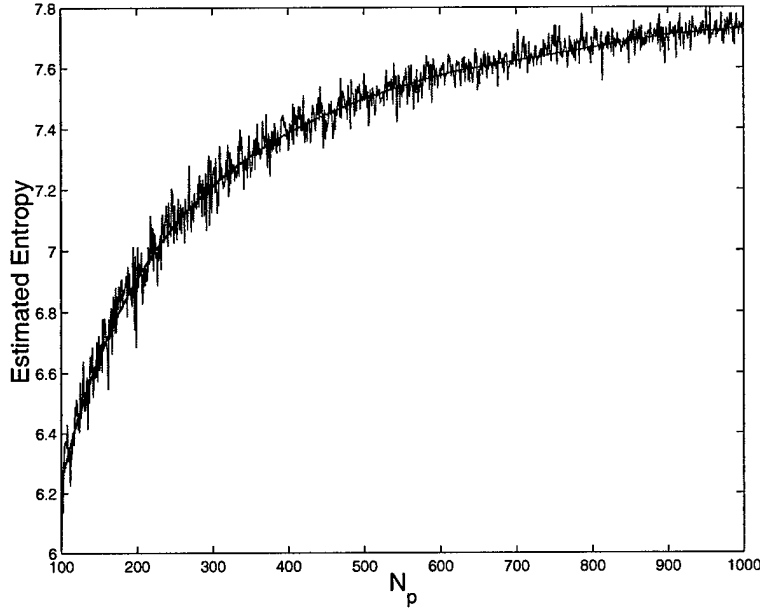


Figure 4.13 Plot demonstrating predictability of N -bias. The solid smooth curve shows the expected entropy estimate, corrected for N -bias, vs. N_p . The jagged curve is the actual experimental entropy estimate. The bias-corrected calculation closely predicts the experimental estimate .

shows the expected N -biased estimate. The jagged curve shows experimental estimates according to (4.27) for different values of N_p . The experimental results closely follow the calculated expected result. Figure 4.14 shows results of a similar simulation. In this case, the upper set of curves are actual and expected entropy estimates when a Gaussian window with $\sigma = 5$ was used to estimate the probabilities. The lower set of curves are those from Figure 4.13 for comparison. The use of the Gaussian window reduces the bias for small N_p , but becomes a slight over-estimate due to the dominance of the S -bias for large N_p .

The variance of the measured entropy estimates without and with the Parzen window technique are more clearly shown in Figures 4.15 and 4.16, respectively. The figures show the difference between the empirically measured entropy, based on N_p samples, and the expected entropy, based on knowledge of the PMF and N_p . The upper and lower smooth curves represent two standard deviations from the expected entropy as calculated using

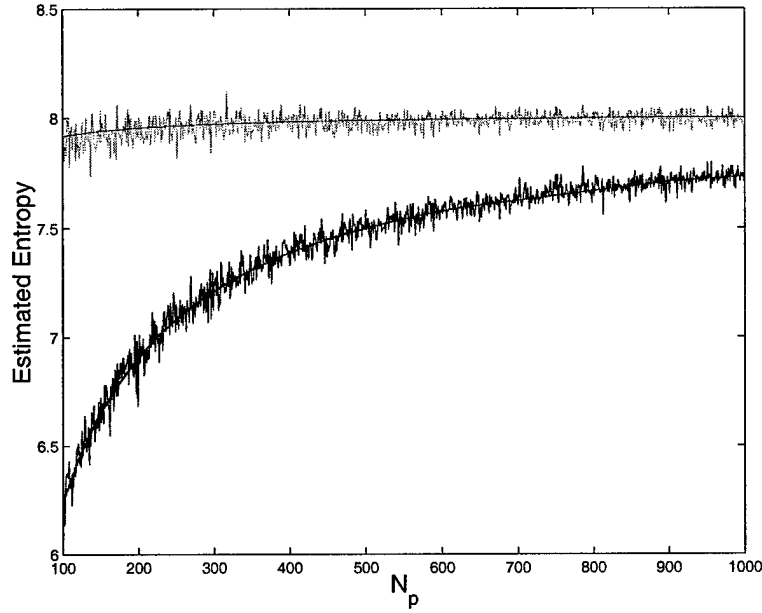


Figure 4.14 Demonstration of the predictability of N -bias and S -bias when Parzen windowing is used. Upper set of curves shows the predicted and actual Parzen-based entropy estimates vs. N_p . Lower set of curves are those of Figure 4.13 for reference.

(4.35). As expected, approximately 95% of the samples lie within the 2σ interval. The magnitudes of the variations are quite similar with and without Parzen smoothing.

As a final note on the bias and variance, experiments were conducted to determine the effect of a Parzen window on the location of the global maximum for mutual information. Figure 4.17 shows the MI vs. shift plot, similar to Figure 4.8, of the 1-D images from Figure 4.4 using a Gaussian window of varying widths. Increasing the window width increases the underestimation of the MI and reduces the dynamic range, but the global minimum does not change position. Figure 4.18 shows the curves of Figure 4.17 after scaling them to a dynamic range of (0,1). The normalized curves are nearly identical for a wide range of windows. Similar results are obtained using a rectangular window with width ranging from between 1 and 120 pixels.

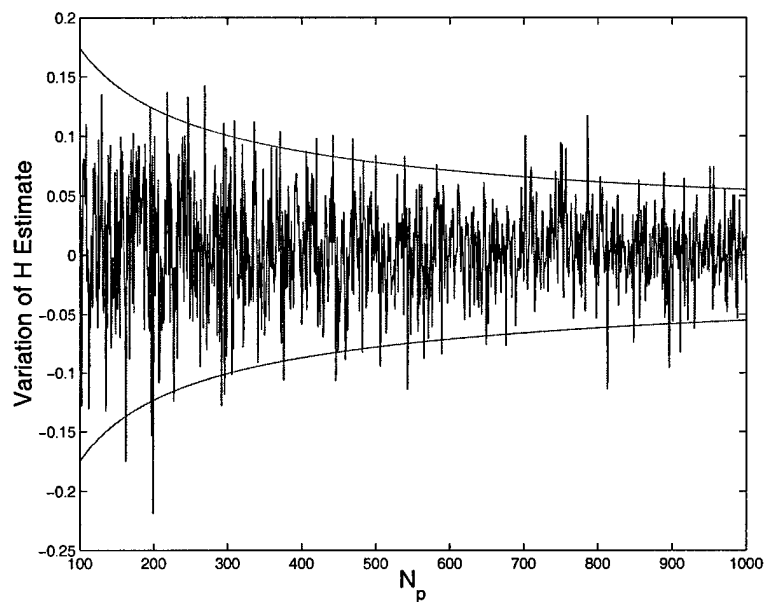


Figure 4.15 Demonstration of the predictability of entropy estimation variance. The upper and lower smooth curves represent two standard deviations from the mean as calculated by (4.35).

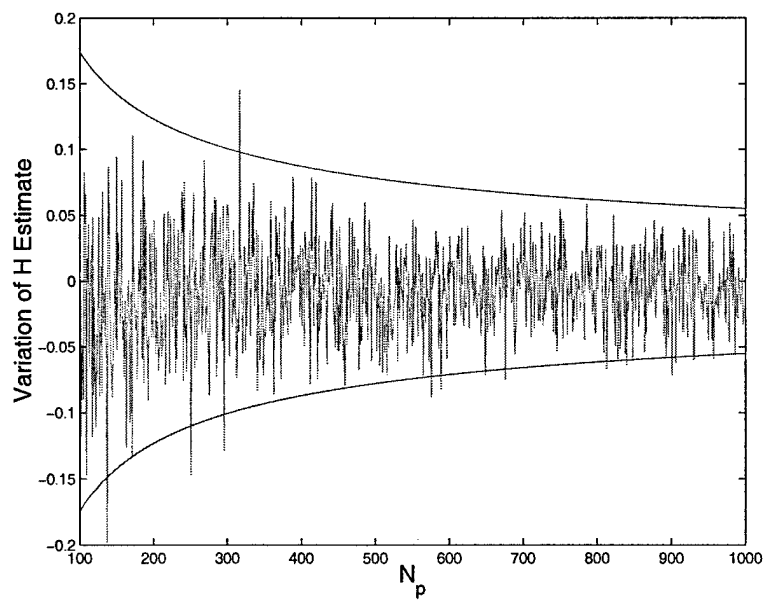


Figure 4.16 Demonstration of the predictability of entropy estimation variance when Parzen windowing is used. The variance is quite similar in magnitude to that of Figure 4.15. The curved lines are added as a reference for comparison with the variance plot of Figure 4.15.

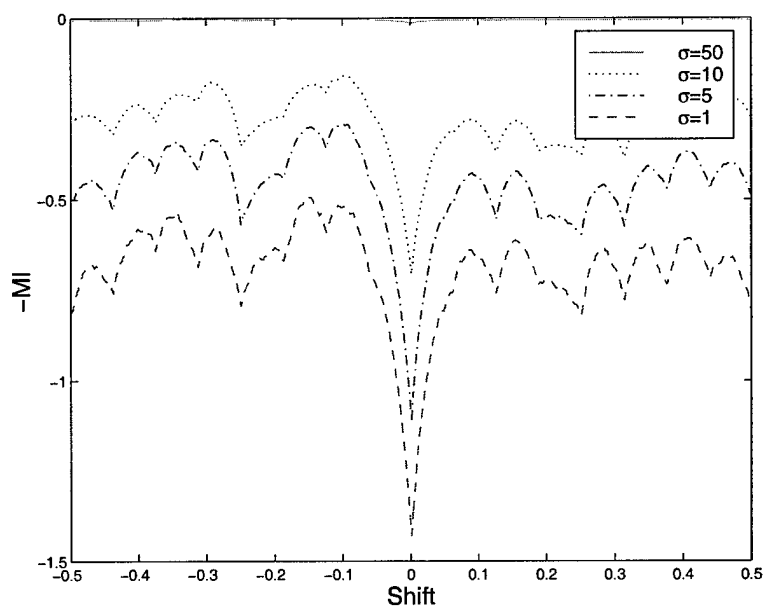


Figure 4.17 Alignment error curves using MI and Gaussian-shaped Parzen windows of varying widths. As the window width increases, the dynamic range of the curve is reduced and the bias increases.

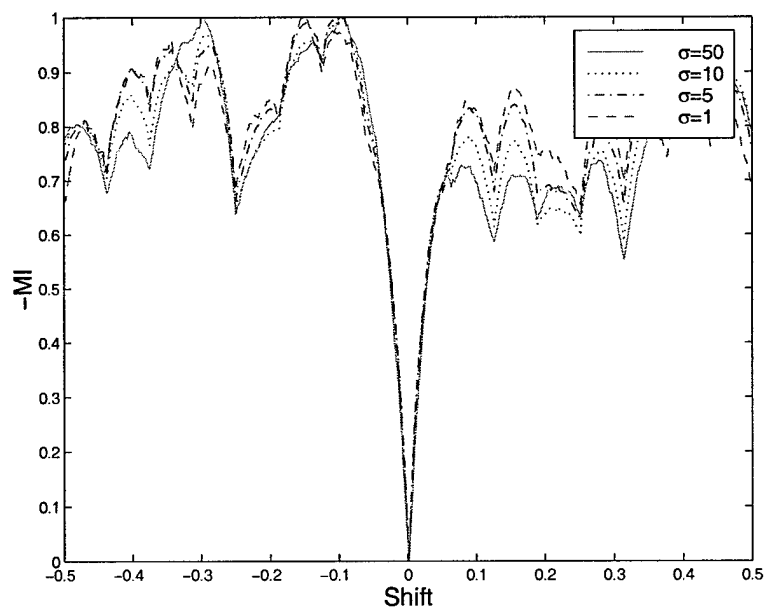


Figure 4.18 Normalized alignment error curves using MI and Gaussian-shaped Parzen windows of varying widths. The misalignment curve is quite similar for a wide range of window widths.

4.4 Bound on the Maximal Mutual Information

An upper bound is now defined for the MI between $\mathbf{d}$ and $\tilde{\mathbf{a}}$. To derive the bound, $\mathbf{u}$ must be limited appropriately. In the context of image registration, $\mathbf{u}$ defines a coordinate transform, so $\tilde{\mathbf{a}}(\mathbf{x}) = \mathbf{a}(\mathbf{x} - \mathbf{u})$ maintains the same number of graylevels as $\mathbf{a}$. Let $G(\tilde{\mathbf{a}})$ be a function determining the number of graylevels in an image, and define the following sets:

$$\omega_u = \{\mathbf{u} \mid G(\mathbf{a}(\mathbf{x} - \mathbf{u})) \leq G(\mathbf{a})\} \quad (4.46)$$

and

$$\omega_a = \{\tilde{\mathbf{a}} \mid G(\tilde{\mathbf{a}}) \leq G(\mathbf{a})\} \quad (4.47)$$

It immediately follows that

$$\max_{\mathbf{u} \in \omega_u} I(\mathbf{d}; \tilde{\mathbf{a}}) = \max_{\mathbf{a} \in \omega_a} I(\mathbf{d}; \mathbf{a}) \triangleq B \quad (4.48)$$

and

$$I(\mathbf{d}; \tilde{\mathbf{a}}) \leq B \quad (4.49)$$

The equality of (4.48) is obvious noting that $\omega_a = \{\tilde{\mathbf{a}} \mid \mathbf{u} \in \omega_u\}$.

A more insightful bound for $I(\mathbf{d}; \tilde{\mathbf{a}})$ is now derived.

Theorem 4.1 Consider two images, $\mathbf{d}$ and $\mathbf{a}$, such that $G(\mathbf{a}) \leq G(\mathbf{d})$. Let $\hat{\omega}_a = \{\mathbf{a} \mid H(\mathbf{a} \mid \mathbf{d}) = 0, \mathbf{a} \in \omega_a\}$; then,

$$B = \max_{\mathbf{a} \in \hat{\omega}_a} H(\mathbf{a}) \quad (4.50)$$

Proof: First, note that for any $\mathbf{a} \in \hat{\omega}_a$

$$I(\mathbf{d}; \mathbf{a}) = H(\mathbf{a}) - H(\mathbf{a} \mid \mathbf{d}) = H(\mathbf{a})$$

Also, $\hat{\omega}_a$ is not empty under the condition $G(\mathbf{a}) \leq G(\mathbf{d})$. This can be understood from the fact that there always exists an $\mathbf{a} \in \hat{\omega}_a$ such that

$$P(a_j \mid d_i) = 0 \text{ or } P(a_j \mid d_i) = 1, \quad \forall i, j \quad (4.51)$$

where d_i and a_i denote the intensity values of $\mathbf{d}$ and $\mathbf{a}$, respectively. For example, (4.51) is achieved when the intensity mapping from $\mathbf{d}$ to $\mathbf{a}$ is one-to-one (in the case of $G(\mathbf{a}) = G(\mathbf{d})$) or many-to-one (if $G(\mathbf{a}) < G(\mathbf{d})$).

Next, (4.50) is proven. Note that $\hat{\omega}_a \subset \omega_a$, so

$$\max_{\mathbf{a} \in \omega_a} I(\mathbf{d}; \mathbf{a}) \geq \max_{\mathbf{a} \in \hat{\omega}_a} I(\mathbf{d}; \mathbf{a})$$

Therefore, it suffices to show that for any $\mathbf{a} \in \omega_a$, there exists an $\hat{\mathbf{a}} \in \hat{\omega}_a$ such that

$$I(\mathbf{d}; \mathbf{a}) \leq I(\mathbf{d}; \hat{\mathbf{a}}) \quad (4.52)$$

Clearly, if $\mathbf{a}$ happens to be in $\hat{\omega}_a$, then $\hat{\mathbf{a}} = \mathbf{a}$ satisfies (4.52). Now, assume that $\mathbf{a} \notin \hat{\omega}_a$; that is, $H(\mathbf{a} | \mathbf{d}) \neq 0$. Then there exists d_1 mapped to (at least) two different values a_1 and a_2 . Define variables $P_{0,0}$ and Δ such that $P_{DA}(d_1, a_1) = P_{0,0} - \Delta$ and $P_{DA}(d_1, a_2) = \Delta$, where $0 \leq \Delta \leq P_{0,0}$. Also define $P_0 = P_A(a = a_1 | \Delta = 0)$ and $P_1 = P_A(a = a_2 | \Delta = 0)$. The value of Δ can be adjusted by selecting different images $\mathbf{a}_\Delta \in \omega_a$. Based on these definitions,

$$\begin{aligned} I(\mathbf{d}; \mathbf{a}_\Delta) &= H(\mathbf{d}) + H(\mathbf{a}_\Delta) - H(\mathbf{d}, \mathbf{a}_\Delta) \\ &= H(\mathbf{d}) - (P_0 - \Delta) \log(P_0 - \Delta) \\ &\quad - (P_1 + \Delta) \log(P_1 + \Delta) - \sum_{a \neq a_1, a \neq a_2} P(a) \log P(a) \\ &\quad + (P_{0,0} - \Delta) \log(P_{0,0} - \Delta) + \Delta \log \Delta \\ &\quad + \sum_{\substack{(d,a) \neq (d_1, a_1) \\ (d,a) \neq (d_1, a_2)}} P(d, a) \log P(d, a) \end{aligned}$$

Treating $I(\mathbf{d}; \mathbf{a}_\Delta)$ as a function of Δ leads to

$$\frac{dI(\mathbf{d}; \mathbf{a}_\Delta)}{d\Delta} = \log \frac{(P_0 - \Delta)\Delta}{(P_1 + \Delta)(P_{0,0} - \Delta)}$$

and

$$\frac{d^2 I(\mathbf{d}; \mathbf{a}_\Delta)}{d\Delta^2} = \frac{1}{P_{0,0} - \Delta} - \frac{1}{P_0 - \Delta} + \frac{1}{\Delta} - \frac{1}{P_1 + \Delta} \quad (4.53)$$

The first term of (4.53) is always larger than the second term, and the third term is larger than the forth, so $I(\mathbf{d}; \mathbf{a}_\Delta)$ is convex with respect to Δ (see [48], p. 31). From this convex property, it follows that $I(\mathbf{d}; \mathbf{a}_\Delta)$ is maximal when $\Delta = P_{0,0}$ or when $\Delta = 0$. This corresponds to either $P(a_1 | d_1) = 0$ or $P(a_2 | d_1) = 0$, respectively. If the resulting maximal $\mathbf{a}_\Delta$ belongs to $\hat{\omega}_a$, then setting $\hat{\mathbf{a}} = \mathbf{a}_\Delta$ achieves the inequality of (4.52). If $\mathbf{a}_\Delta \notin \hat{\omega}_a$, then following the same analysis, there exist $\tilde{a}_1$ and $\tilde{a}_2$ in $\mathbf{a}_\Delta$ such that $P(\tilde{a}_1 | \tilde{d}_1) \neq 0$ and $P(\tilde{a}_2 | \tilde{d}_1) \neq 0$ for some $\tilde{d}_1$. However, a new $\mathbf{a}_\Delta \in \omega_a$ can be found which has higher mutual information with $\mathbf{d}$, and $P(\tilde{a}_1 | \tilde{d}_1) = 0$ or $P(\tilde{a}_2 | \tilde{d}_1) = 0$. Therefore, a necessary condition on any template $\mathbf{a}^*$ such that $I(\mathbf{d}; \mathbf{a}^*)$ achieves its maximal value is that there exist no a_1 and a_2 such that for a fixed value d_1 both $P(a_1 | d_1) \neq 0$ and $P(a_2 | d_1) \neq 0$. In other words, $\mathbf{a}^* \in \hat{\omega}_a$. This proves (4.52) and the theorem.

Remark 4.1:

$$B \leq \log(G(\mathbf{a})) \quad (4.54)$$

The above inequality can be easily proven noting that

$$\max_{\mathbf{a} \in \hat{\omega}_a} H(\mathbf{a}) \leq \max_{\mathbf{a} \in \omega_a} H(\mathbf{a}) = \log G(\mathbf{a})$$

Remark 4.2:

Equation (4.50) still holds if the set ω_a is tightened such that

$$\omega_a = \{\tilde{\mathbf{a}} \mid G(\tilde{\mathbf{a}}) = G(\mathbf{a})\}$$

This can be understood from the following facts:

1. If $G(\tilde{\mathbf{a}}) < G(\mathbf{a})$ and $\tilde{\mathbf{a}} \in \hat{\omega}_a$, a corresponding $\tilde{\mathbf{a}}'$ can be found such that $G(\tilde{\mathbf{a}}') = G(\mathbf{a})$ and $\tilde{\mathbf{a}}' \in \hat{\omega}_a$.
2. $\max_{\mathbf{a} \in \omega_a} H(\mathbf{a})$ increases as $G(\mathbf{a})$ increases.

4.4.1 Discussion of the bound

It is clear from Theorem 4.1 that the MI bound can be calculated directly by maximizing the entropy of the template under the constraint that the conditional entropy of the

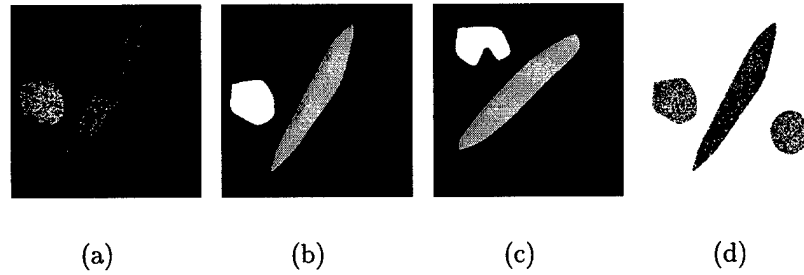


Figure 4.19 Synthetic image with multiplicative noise and several possible segmentations, including one that achieves the bound. (a) Synthetic image with multiplicative noise, (b) correct manual segmentation, (c) an incorrect segmentation, (d) another incorrect segmentation which achieves the bound set by (4.50).

template, given the fixed image, be zero. If all possible transformations of a template are allowed, the bound is obtainable, and the bound-achieving template is not unique. This implies that if a transform is too “powerful” or “flexible,” maximizing the MI may result in an undesirable deformation. To illustrate this point, consider the images in Figure 4.19. Figure 4.19(a) shows an image which has been corrupted by multiplicative noise such that it contains 21 distinct intensities. Figures 4.19(b)-(d) represent candidate template images which all contain four distinct symbols. The bound computed from (4.50) for Figure 4.19(a) and any template containing four symbols is 0.963. The MI between Figures 4.19(a) and 4.19(b) is 0.658, between Figures 4.19(a) and 4.19(c) the MI is 0.110, and Figures 4.19(a) and 4.19(d) achieve the bound.

A similar study was conducted with a cryosection image of a human head. Figure 4.20(a) is the head image and Figures 4.20(b) and (c) are templates each containing five distinct symbols. The MI between Figures 4.20(a) and 4.20(b) is 1.210 and Figures 4.20(a) and 4.20(c) achieve the bound of 1.674.

Although human observers consider Figures 4.19(b) and 4.20(b) to be most closely aligned with Figures 4.19(a) and 4.20(a), Pairs 4.19(a) and 4.19(d) and 4.20(a) and 4.20(c) have the highest MI. Figure 4.21 shows more explicitly the lack of structure in Figure 4.20(c). To avoid this type of problem in deformable image registration, structural

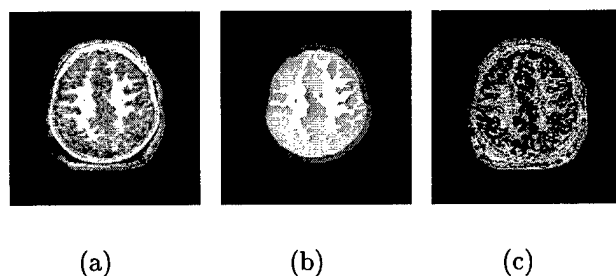


Figure 4.20 Real cryosection image and several possible segmentations, including one that achieves the bound. (a) Cryosection image of a human brain, (b) correct manual segmentation, (c) another segmentation which achieves the bound set by (4.50).

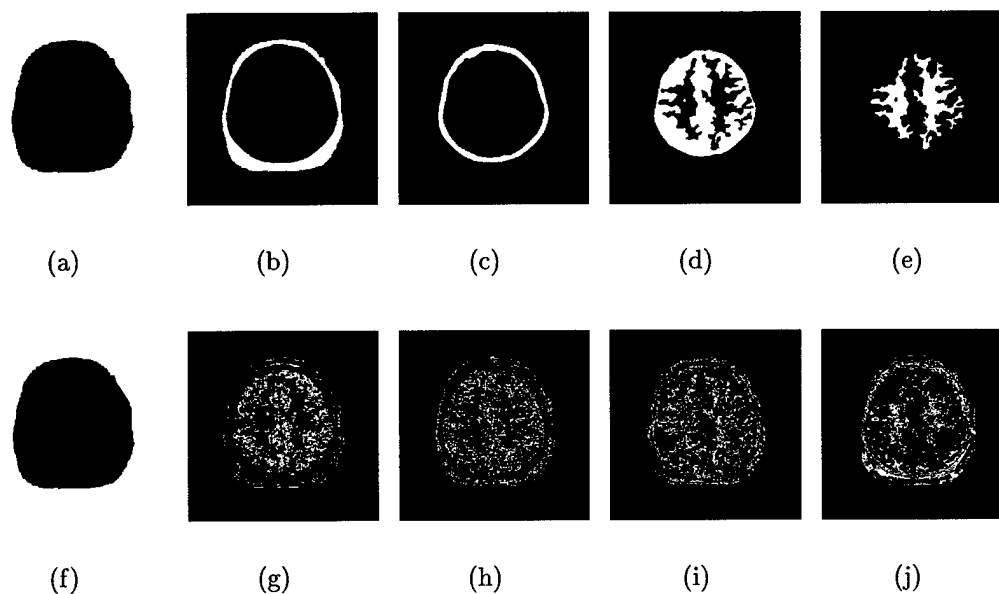


Figure 4.21 Bound achieving template separated by graylevel. Top row shows individual graylevels of Figure 4.20(b). Bottom row shows individual graylevels of Figure 4.20(d). The bottom row clearly lacks the structure of the top row. The graylevel indicating background is the only one which maintains a correct structure.

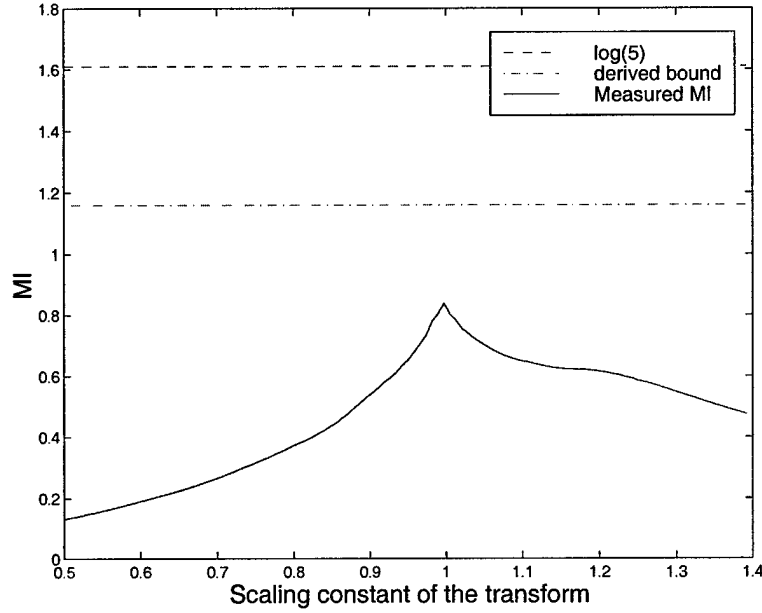


Figure 4.22 Relationship between $\log(G(\mathbf{a}))$, the derived bound, and $I(\mathbf{d}; \tilde{\mathbf{a}})$, where $\mathbf{u}$ is a scaling transform. The curves are based on the data in Fig. 4.20.

and topological constraints should be used, as is the case with elastic or fluid models [34, 57]

Figure 4.22 illustrates the relationship between the bounds in (4.50) and (4.54) and $I(\mathbf{d}; \tilde{\mathbf{a}})$ based on the data in Figure 4.20. Here we assume $\mathbf{u}$ is a simple scaling transform. As can be seen, the derived bound is much tighter than $\log(G(\mathbf{a}))$. An even tighter bound could also be derived, but it would be dependent on the characteristics of the transform and the image data.

4.5 Optimality of Mutual Information

From the discussion of Section 4.4.1 it is clear that, in some cases, the template which maximizes $I(D; A)$ will not correspond to the correct structure. It is instructive to determine the circumstances under which the maximal MI will give meaningful results. It can be proven that if the imaging function C is one-to-one, then MI is optimal in the sense that the correct segmentation is expected to maximize the MI. Recall from (4.16) that maximizing $I(D; \tilde{A})$ is equivalent to minimizing $H(D | \tilde{A})$, and let $f(\cdot)$ be a

one-to-one mapping; then,

$$H(f(s) + \eta \mid s) = H(\eta) \quad (4.55)$$

$$= H(\eta \mid \tilde{a}) \quad (4.56)$$

$$\leq H(f(s) + \eta \mid \tilde{a}), \quad \forall \tilde{a} \quad (4.57)$$

In other words, the MI between the correct segmentation and the study image is expected to be at least as large as the MI for any other segmentation choice. If, however, the graylevel mapping is not one-to-one, the equality of (4.55) may not hold, so s may not be the maximal MI choice.

It should also be noted that (4.55) and (4.57) assume white noise. In reality any particular instance of the white noise process will be correlated. In fact, it is the goal of the maximal MI deformation to correlate the template to the image as much as possible. So, although the *expected* maximizing deformed template is s , the deviation from the expected template for any particular study image will depend on both the noise and the structure of the image.

There is ambiguity associated with the equality portion of (4.57). It is manifest in the fact that symmetries are indistinguishable. The three images of Figure 4.23 are considered equivalent using the MI metric. This may not pose a serious threat, since the segmentations boundaries are not changed.

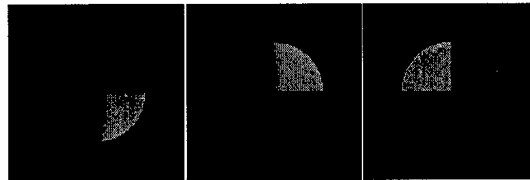


Figure 4.23 Symmetry problem example. The three images are equivalent in an information sense.

Figure 4.24 illustrates one way a non-one-to-one mapping can lead to problems. Figure 4.24(a) has a structure most similar to Figure 4.24(b), but because Figure 4.24(b) uses the black intensity for two different structures, the MI between (a) and (c) is greater

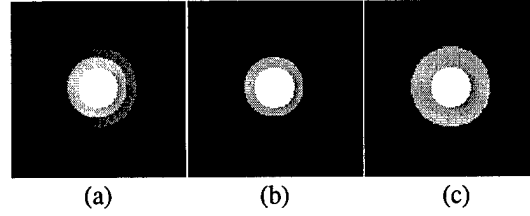


Figure 4.24 Many-to-one graylevel mapping problem example. Visually, (a) and (b) have equivalent structure, but the mutual information is highest for the pair (a) and (c).

than (a) and (b). This problem can be partially avoided by selecting a template such that each structure uses a different graylevel.

4.5.1 Maximum likelihood optimality

The fundamental assumption governing the use of mutual information is that the mutual information between the study image $\mathbf{d}$ and the deformed template $\tilde{\mathbf{a}}$ is maximal when the template is equivalent to the underlying structure $\mathbf{s}$ of the image. Under certain conditions, this is equivalent to the maximum likelihood decision. In a probabilistic sense, an optimal template can be defined by

$$\hat{\mathbf{s}} = \arg \max_{\tilde{\mathbf{a}}} P(\tilde{\mathbf{a}} \mid \mathbf{d}) \quad (4.58)$$

Using the definition of information (3.53) and the monotonicity of the log function, (4.58) becomes

$$\hat{\mathbf{s}} = \arg \min_{\tilde{\mathbf{a}}} I(\tilde{\mathbf{a}} \mid \mathbf{d}) \quad (4.59)$$

$$= \arg \min_{\tilde{\mathbf{a}}} I(\mathbf{d} \mid \tilde{\mathbf{a}}) + I(\tilde{\mathbf{a}}) \quad (4.60)$$

where the second equality follows from the information theory equivalent of Bayes' rule. In other words, when determining which deformed template is best aligned with the underlying structure of an image, select the deformed template which is least "surprising" based on the observed data.

An information-theoretic approach is used for three reasons:

- It provides a theoretical basis through which optimization can be achieved independent of the amount of prior knowledge about the imaging system. In the absence of information about the system, a solution can be found which minimizes the complexity of the system.
- The problem can be related to concepts in information theory — such as information, mutual information, and entropy — which aid in the understanding of the problem.
- Mathematical identities and theorems are in place which facilitate analysis.

The second term on the right-hand side of 4.60 is the prior information about the template. Minimization of this term will be discussed in Chapters 5 and 6; the main focus here is on the first term.

A major problem with applying (4.60) in its current form is that $P(\mathbf{d}|\tilde{\mathbf{a}})$ is difficult to estimate since we only have one example of $\mathbf{d}$. However, assuming elements of the random vector $(\mathbf{d}|\tilde{\mathbf{a}})$ are independent and identically distributed (equivalently, the imaging process is memoryless and stationary), the first term of (4.60) can be rewritten in a more useful form. From the identical distribution assumption define variables d and $\tilde{a}$ such that

$$I(\mathbf{d}(i) | \tilde{\mathbf{a}}(i)) = I(\mathbf{d}(j) | \tilde{\mathbf{a}}(j)) = I(d | \tilde{a}), \quad \forall i, j \quad (4.61)$$

Then (neglecting for a moment the prior information),

$$\hat{s} = \min_{\tilde{\mathbf{a}}} I(\mathbf{d} | \tilde{\mathbf{a}}) \quad (4.62)$$

$$= \min_{\tilde{\mathbf{a}}} [-\log P(\mathbf{d} | \tilde{\mathbf{a}})] \quad (4.63)$$

$$= \min_{\tilde{\mathbf{a}}} \sum_{i=1}^{N_p} -\log P(\mathbf{d}(i) | \tilde{\mathbf{a}}(i)) \quad (4.64)$$

$$= \min_{\tilde{\mathbf{a}}} - \sum_{\tilde{a}=0}^{N_a-1} \sum_{d=0}^{N_d-1} P_{D\tilde{A}}(d, \tilde{a}) \log P_{D|\tilde{A}}(d | \tilde{a}) \quad (4.65)$$

$$= \min_{\tilde{\mathbf{a}}} H(D | \tilde{A}) \quad (4.66)$$

$$= \max_{\tilde{\mathbf{a}}} I(D; \tilde{A}) \quad (4.67)$$

Line (4.63) follows from the definition of information in (3.53), and (4.64) is a result of the independence assumption. Line (4.65) follows from (4.61) and the fact that $P_{D\tilde{A}}(d, \tilde{a})$ is directly proportional to the number of times graylevels d and $\tilde{a}$ occur concurrently in the image and template, respectively. Thus, the summation over all pixel locations is converted to a summation over all graylevels in both the image and template.

One conclusion of (4.62)-(4.67) is that maximization of mutual information is optimal in a Bayesian sense under certain conditions. The conditions are independent, identically distributed elements of $(\mathbf{d} \mid \tilde{\mathbf{a}})$ and equally likely $\tilde{\mathbf{a}}$. As previously mentioned in Section 2.4, several information-based metrics have been proposed as alignment metrics. Studies have shown that each of them may perform better than others under specific circumstances; however, MI has been most widely accepted for general application. This is consistent with the probabilistic optimality of MI for unknown contrast.

A second conclusion is that the problem of finding the least surprising deformed template considering only the image data is equivalent to minimizing $H(D \mid \tilde{A})$, which is the uncertainty introduced by the imaging system; this is equivalent to finding the simplest imaging process. For example, a very simple channel (simple in the Kolmogorov sense, but maybe hard to construct!) may just produce s as the output, but a very complex imaging channel might make a battleship look more like a butterfly. One may not assume, however, that the imaging process used in any instance is a minimum entropy process. The result of this assumption may lead to images similar to those of Figures 4.19(d) and 4.20(d).

It is notable that, in going from (4.64) to (4.65), the minimization changes from dependence on N_p to dependence on N_a and N_d . Thus, the MI calculation is a constant with respect to the image size.

4.6 Summary

This chapter has investigated use of mutual information as an alignment metric for non-rigid-body template matching. It was shown that MI has the desirable characteristic

that knowledge of the imaging function is not required. Thus it is useful for images from a wide range of imaging methods. An in-depth study was conducted on the bias and variance of the empirical MI, as well as the use of Parzen windows to estimate the probabilities associated with the images. It was argued theoretically that Parzen window estimation may reduce sensitivity of a deformation to noise and partial volume effects.

A gradient calculation was presented for use in the deformation process. The calculation scales linearly with the image size, so it is readily scalable to large images. Utility of the gradient descent method was found to be dependent on the image and template structures. The derived gradient method is valid for both normalized histogram-based MI and Parzen window-based MI.

A bound was also found on the maximal MI between a given image and a template. The bound provided useful insight into the use of MI as an alignment metric for template deformation.

Finally, the conditions for optimality were explained under two different optimality criteria. The conditions for Bayesian optimality are as follows:

- Elements of the random vector $\mathbf{d}$ are independent given $\tilde{\mathbf{a}}$.
- Elements of $(\mathbf{d}|\tilde{\mathbf{a}})$ are identically distributed.
- The possible $\tilde{\mathbf{a}}$ have equally likely probability.

A sufficient condition to ensure that the true underlying structure is the maximal MI structure is that the imaging process be one-to-one. Problems related to non-one-to-one imaging processes can be mitigated by reducing the number of graylevels in the template and using a separate graylevel for each structure.

CHAPTER 5

A FLUID-BASED TRANSFORM MODEL

This chapter addresses the question of how to determine a displacement field so as to maximize the similarity metric while still maintaining a meaningful result. To accomplish this, a transformation model $\mathcal{L}$ is defined, which models all the feasible transformations. The displacement field is related to the body force (and thus the study image) through

$$\mathcal{L}\mathbf{u} + \mathbf{b} = 0 \quad (5.1)$$

where $\mathbf{b}$ is the body force and $\mathbf{u}$ is the template displacement. For image segmentation, it is desirable that the deformation preserve the template topology and accommodate large, curved displacements. It has been shown by Christensen et al. [25, 61] that deformations based on a non-mass-conserving viscous fluid model preserve these features. This chapter first reviews this fluid model, then proposes a modified model that preserves the desirable properties of the fluid model, but requires less computation and may lead to more intuitive incremental deformations.

5.1 Terminology and Relevant Theorems

5.1.1 Topology conservation

Two images are considered *topologically equivalent* if and only if the two images can be related through a homeomorphism. A *homeomorphism* is a continuous, one-to-one,

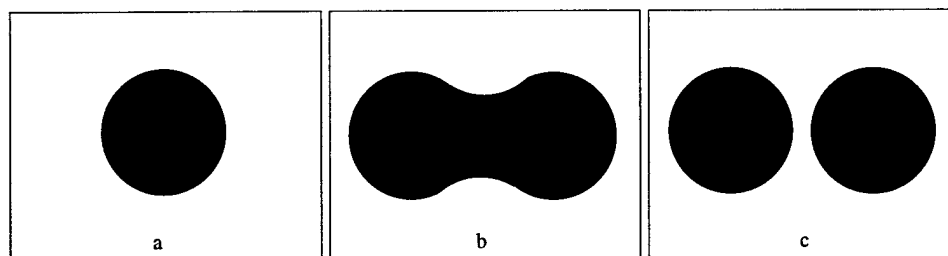


Figure 5.1 Homeomorphic vs. nonhomeomorphic. (a) and (b) are homeomorphic, but (c) is not homeomorphic with either (a) or (b). The single circle of (a) or (b) would have to be torn to deform to (c).

onto mapping between two geometric figures, which also has a continuous inverse. Homeomorphic mappings include operations such as bending and stretching; tearing, however, is not allowed. A sphere, for example, is topologically equivalent to an ellipsoid because they are related through stretching. A single sphere is not equivalent to two spheres since the single sphere would have to be torn to map to two spheres. This is illustrated in Figure 5.1. In the case of a deformation of Figure 5.1(a) to 5.1(c), the point at which white space appears between the two circles in (c) marks a discontinuity in the mapping between the images since it is ambiguous whether the particle comes from above or below the initial sphere.

5.1.2 Large, nonlinear deformations

Deformable models such as linear elastic models assume that deformations are small enough that the nonlinear terms present in the actual physical elastic model may be neglected. Deformations of this type follow straight-line trajectories, and are penalized in proportion to the distance of the deformation.

On the other hand, large deformation models, such as the fluid model used in this thesis, do not expressly restrict the magnitude of the displacements. Instead, only the rate of displacement is regulated. The model is also nonlinear because the displacements are allowed to follow curved trajectories.

5.1.3 Eulerian formulation

Deformations are represented using an Eulerian formulation. Within continuum mechanics, the coordinate system used to formulate problems is based on either an Eulerian or Lagrangian formulation. Eulerian formulations define a fixed coordinate system, x , y , z , and t . The fluid at a specific location in the coordinate system will consist of different actual fluid at different times, based on the flow of the fluid. In contrast, under a Lagrangian formulation the coordinate system moves with a particular volume of fluid being tracked. In this thesis, the Eulerian formulation is used. The Eulerian formulation usually leads to smaller errors from interpolation effects. The Eulerian formulation is also more convenient when comparing the template to a study image and for display purposes. As the template deforms, moving (Lagrangian) coordinates may not correspond to the fixed (Eulerian) coordinates used in the study image. Thus, an extra interpolation step would be required to convert the relocated points to the values on the discrete lattice.

Figure 5.2 illustrates the relationship between the deformed template, the original template, and the displacement field. The displacement vector field at any point $\mathbf{x}$ and time t points from the original particle's location to its current location. Thus, the particle at $\mathbf{x}$ at time t originated at location $\mathbf{x} - \mathbf{u}(\mathbf{x})$ in the original template at time $t = 0$.

5.1.4 Pertinent theorems

As the transform described in this chapter is a homeomorphism, it is instructive to present some definitions and theorems by which this property can be proven. Proofs of the theorems have been omitted.

A transformation is said to be continuous if it satisfies the following definition.

Definition 5.1 ([62], p. 332) *A transformation $\mathbf{h}$ defined on a set $D \subset \mathbb{R}^n$ is said to be continuous at point $\mathbf{p}_0 \in D$ if and only if for any $\epsilon > 0$ there is a $\delta > 0$ such that $|\mathbf{h}(\mathbf{p}) - \mathbf{h}(\mathbf{p}_0)| < \epsilon$ whenever $|\mathbf{p} - \mathbf{p}_0| < \delta$ and $p \in D$.*

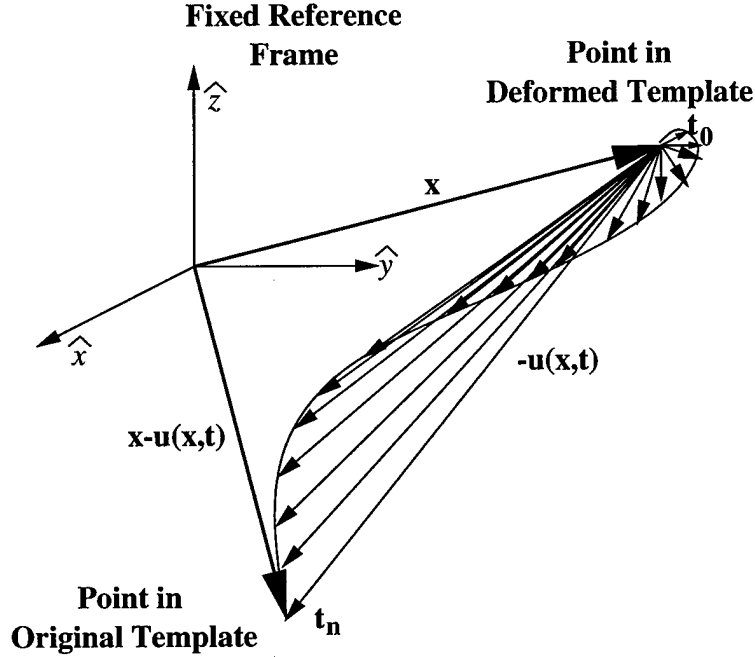


Figure 5.2 Illustration of the relationship between a point fixed point in the deformed template and points in the original template as time varies. (Adapted from [25].)

Furthermore, a transformation that is continuous for every point in D is said to be continuous on D .

The next definition and theorem state that for continuous transforms, structures that are adjacent to each other before transformation are still adjacent after transformation even though they may have changed shape.

Definition 5.2 ([63]) *A connected set is a set which cannot be partitioned into two nonempty subsets which are open in the relative topology induced on the set. Equivalently, a connected set cannot be partitioned into two nonempty subsets such that each subset has no points in common with the closure of the other.*

Theorem 5.1 ([62], p. 333) *Let $\mathbf{h}$ be continuous on a set D . Then any connected set $S \subset D$ is carried into a connected set $\mathbf{h}(S)$.*

Also, any concatenation of continuous transforms is a continuous transform, as is stated by the next theorem.

Theorem 5.2 ([62], p. 333) *Let the transformation $\mathbf{g}$ be continuous on a set $\mathbb{A}$ and $\mathbf{h}$ be continuous on a set $\mathbb{B}$, and let $\mathbf{p}_0 \in \mathbb{A}$ and $\mathbf{g}(\mathbf{p}_0) \in \mathbb{B}$. The product transformation $\mathbf{h}(\mathbf{g}(\mathbf{p}))$ is continuous at $\mathbf{p}_0$.*

5.2 Fluid Model

In the viscous fluid model presented by Christensen et al. [25, 61], a transform $\mathbf{h}(\mathbf{x}, t)$ is computed by

$$\mathbf{h}(\mathbf{x}, t) = \mathbf{x} - \mathbf{u}(\mathbf{x}, t) \quad (5.2)$$

where $\mathbf{x}$ takes on values in $\mathbb{X}$. The notion of a velocity field $\mathbf{v}(\mathbf{x}, t)$ is introduced which is related to the displacement field through the full derivative of $\mathbf{u}$ by

$$\mathbf{v}(\mathbf{x}, t) = \frac{d}{dt}\mathbf{u}(\mathbf{x}, t) \quad (5.3)$$

The solution to this ordinary differential equation (ODE) is

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{u}(\mathbf{x}, 0) + \int_0^t \mathbf{v}(\mathbf{x}, \tau) d\tau \quad (5.4)$$

with conditions on the boundary $\partial\mathbb{X}$ of $\mathbb{X}$ that $\mathbf{v}$, and hence $\mathbf{u}$ and $\mathbf{h}$, equal zero.

To remain within the Eulerian framework, the full or *material derivative* is related to the partial derivative of $\mathbf{u}$ through a material derivative expansion

$$\mathbf{v}(\mathbf{x}, t) = \frac{\partial}{\partial t}\mathbf{u}(\mathbf{x}, t) + \nabla[\mathbf{u}(\mathbf{x}, t)]^T \mathbf{v}(\mathbf{x}, t) \quad (5.5)$$

The velocity is chosen to mimic that of a viscous fluid in response to a force proportional to the steepest ascent direction of a similarity metric:

$$\mu_1 \nabla^2 \mathbf{v} + \mu_3 \nabla \nabla \cdot \mathbf{v} - \mu_2 \mathbf{v} + \mathbf{b} = 0 \quad (5.6)$$

along with boundary conditions $\mathbf{v}|_{\partial\mathbb{X}} = 0$. Equations (5.5) and (5.6) comprise the fluid model, and henceforth (5.6) will be referred to as the “original” partial differential equation (PDE) for the viscous fluid model. The model is $\mathcal{L} = \mathcal{L}_2 \circ \mathcal{L}_1$, where $\mathcal{L}_2 = \mu_1 \nabla^2 + \mu_3 \nabla \nabla \cdot - \mu_2$, $\mathcal{L}_1 = \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla$, and $\circ$ denotes the composition of $\mathcal{L}_2$ and $\mathcal{L}_1$.

5.3 Discrete Approximation to the Fluid Model

The displacement field of (5.3), when approximated in the time domain by finite-differences, is quite easily translated into an explicit formula for the displacement. Based on knowledge of the velocity and displacement at the previous time t^k , the displacement at time t^{k+1} is

$$\begin{aligned} \mathbf{u}(\mathbf{x}, t^{k+1}) \approx & \mathbf{u}(\mathbf{x}, t^k) + \Delta [\mathbf{v}(\mathbf{x}, t^k) \\ & - \nabla[\mathbf{u}(\mathbf{x}, t^k)]^T \mathbf{v}(\mathbf{x}, t^k)] \end{aligned} \quad (5.7)$$

where Δ is a small time increment.

This move to a discrete approximation to the solution of (5.3) may lead to singularities in the transform. In order to guarantee the preservation of topology, the discrete Jacobian of the transformation is determined at each time step. The template is *regridded* [25] if it is found that the transformation is nearly singular at any point in the domain. In the current implementation, a new template is propagated when the Jacobian at any point in the transform drops below a value of 0.5. A larger value causes more templates to be propagated, which will increase numerical precision errors of the transformation due to the increased number of concatenated transformations; conversely, a smaller propagation threshold causes less templates to be propagated. This may increase the numerical precision errors due to interpolating transformations that are nearly singular in places.

The algorithm for a discrete template deformation using the fluid model is as follows:

1. Let $t=0$, $i=0$, $\tilde{\mathbf{a}}^{(0)}(\mathbf{x}) = \mathbf{a}(\mathbf{x})$, and $\mathbf{u}^{(0)}(\mathbf{x}) = 0$.
2. Calculate the body force $\mathbf{b}(\mathbf{x}, t)$ as in Section 4.2.
3. If $\mathbf{b}$ is below a threshold for all $\mathbf{x} \in \mathbb{X}$, or if the maximum number of iterations is reached, then STOP. Otherwise, proceed to step 4.
4. Solve the PDE of (5.6) to find $\mathbf{v}$.
5. Calculate the incremental perturbation of the displacement field by solving for $\frac{\partial \mathbf{u}}{\partial t}$ in (5.5) based on $\mathbf{r} = \mathbf{v}(\mathbf{x}, t) - \nabla[\mathbf{u}(\mathbf{x}, t^k)]^T \mathbf{v}(\mathbf{x}, t^k)$.

6. Choose a real-time step size Δ which is a function of $\max_{\mathbf{x}} |\mathbf{v}(\mathbf{x})|$.
7. If the Jacobian of $\mathbf{x} - \mathbf{u} - \Delta \mathbf{r}(\mathbf{x})$ is less than 0.5, propagate to template $i + 1$ using $\mathbf{a}^{(i+1)}(\mathbf{x}, t) = \mathbf{a}^{(i)}(\mathbf{x} - \mathbf{u}^{(i)}(\mathbf{x}, t))$, $\mathbf{u}^{(i+1)}(\mathbf{x}, t) = 0$, set $i = i + 1$, and go to step 5. Otherwise, update the i th displacement field using $\mathbf{u}^{(i)}(\mathbf{x}, t + \Delta) = \mathbf{u}^{(i)}(\mathbf{x}, t) + \Delta \mathbf{r}(\mathbf{x})$, set $t = t + \Delta$, and go to step 2.

5.4 Modified Fluid Model

Two modifications are made to the fluid model. The first one can be thought of as a judicious choice of μ_3 which leads to a more intuitive deformation in many cases and also a significant reduction in the computational complexity. Second, the complexity is further reduced by careful selection of boundary conditions.

Insight about the fluid model is gained by realizing that (5.6), using discrete time steps, corresponds to rendering stationary the functional [64]

$$F(\mathbf{v}) = \frac{1}{2} \int_{\mathcal{X}} [\mu_1 |\nabla \mathbf{v}|^2 + \mu_3 |\nabla \cdot \mathbf{v}|^2 + \mu_2 |\mathbf{v}|^2 - 2\mathbf{v} \cdot \mathbf{b}] d\mathbf{x}^3 \quad (5.8)$$

The first term penalizes changes in the velocity field, the second term penalizes diverging velocities, and the third term penalizes large velocities. The final term penalizes misalignment between study and template; for the purposes of this thesis, $\mathbf{b}$ is a function of the MI between the images as explained in Section 4.2.

Regularization of the divergence of the velocity field in (5.8) may lead to deformations which are orthogonal to the applied force. An example of the orthogonal displacements is shown in Figure 5.3. A force as applied in Figure 5.3(a) may produce a deformation as shown in Figure 5.3(b). For different values of μ_1 , μ_2 , and μ_3 , however, the same force could produce a different deformation as shown in Figure 5.3(c). Finally, for some shapes, this orthogonal coupling is undesirable all together. For example suppose it is desired to deform the square of Figure 5.4(a) to a rectangle shape. Forces applied as shown in Figure 5.4(a) would result in a deformation shown in Figure 5.4(b).

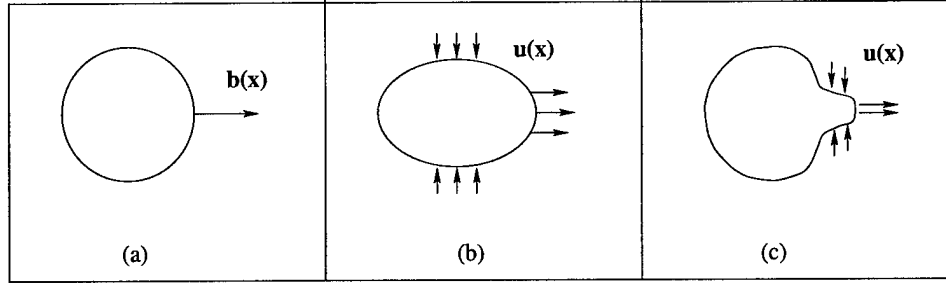


Figure 5.3 Demonstration of orthogonally induced displacements. Applying a force in one direction, (a), may introduce orthogonal displacements as in (b) or (c).

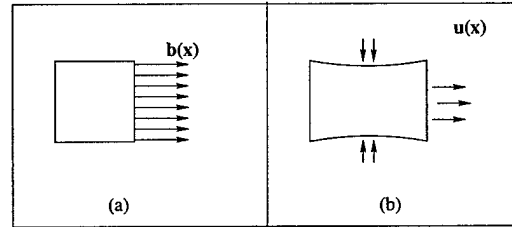


Figure 5.4 A second demonstration of orthogonally induced displacements. Applying a force in one direction, (a), may introduce orthogonal displacements as in (b).

The amount of orthogonal coupling is regulated by μ_3 . Interestingly, although the incremental deformations may be noticeably different for different values of μ_3 , experience has shown the final deformed result to be quite insensitive to a wide range of μ_3 . Letting $\mu_3 = 0$, a much simpler PDE emerges:

$$\mu_1 \nabla^2 \mathbf{v} - \mu_2 \mathbf{v} + \mathbf{b} = 0 \quad (5.9)$$

which is now actually a set of three *uncoupled* scalar PDEs. The numerical solution of (5.9) is generally much faster than that of (5.6) because the number of simultaneous unknowns being found is reduced by a factor of three. Equation (5.9) is referred to as the “modified” PDE.

In order to further render (5.9) amenable to numerical analysis, consider the *extension* of the image and template from the unit cube to all of $\mathbb{R}^3$. Additionally, consider the boundary condition on the velocity field $\lim_{\mathbf{x} \rightarrow \infty} |\mathbf{v}(\mathbf{x})| = 0$. In this case, the response of the velocity field to a force is *translationally invariant*, meaning that the response is a function of the distance between the applied force and response only, not the absolute

location. In this case, the velocity field may be explicitly expressed as

$$\mathbf{v}(\mathbf{x}) = \int_{\mathcal{X}} \bar{\mathbf{G}}(\mathbf{x} - \mathbf{x}') \cdot \mathbf{b}(\mathbf{x}') d\mathbf{x}' \quad (5.10)$$

where the *dyadic Green's function* (impulse response) is given by

$$\bar{\mathbf{G}}(\mathbf{x} - \mathbf{x}') = -\bar{\mathbf{I}} \frac{1}{2\pi\mu_1} \exp\left(-\sqrt{\frac{\mu_2}{\mu_1}}|\mathbf{x} - \mathbf{x}'|\right) \log|\mathbf{x} - \mathbf{x}'| \quad (5.11)$$

In the above, $\bar{\mathbf{I}}$ is the 2-D identity tensor of rank 2. For three-dimensional data, the dyadic Green's function would be

$$\bar{\mathbf{G}}(\mathbf{x} - \mathbf{x}') = \bar{\mathbf{I}} \frac{1}{2\pi\mu_1} \frac{\exp\left(-\sqrt{\frac{\mu_2}{\mu_1}}|\mathbf{x} - \mathbf{x}'|\right)}{|\mathbf{x} - \mathbf{x}'|} \quad (5.12)$$

where $\bar{\mathbf{I}}$ is the 3-D identity tensor of rank 2.

Because (5.11) is nothing more than a convolution, it may be evaluated in the spectral domain as a contraction:

$$\mathbf{v}(\mathbf{x}) = \mathcal{F}^{-1}(\mathcal{F}(\bar{\mathbf{G}}(\mathbf{x})) \cdot \mathcal{F}(\mathbf{b}(\mathbf{x}))) \quad (5.13)$$

where $\mathcal{F}(\cdot)$ is the spatial Fourier transformation operator. Finding the velocity field by evaluating the Fourier integrals in (5.13) with fast transforms has important advantages over directly solving the PDEs in (5.6) and (5.9). Because of the rather large number of unknowns associated with the solution to these PDEs for image resolutions of practical interest, it is necessary to resort to *iterative* solution methods. The most popular of these is perhaps successive over-relaxation (SOR) as used in [25], which requires careful selection of a relaxation parameter for efficient performance. The spectral approach embodied in (5.13) is parameter-free, requires no tuning, and has computational complexity which scales more favorably with image size than that of the SOR [65].

The resulting transform model produces transformations which are homeomorphisms. The proof is similar in nature to a proof for diffeomorphisms in [47]. The steps are as follows: (1) Show that $\mathbf{v}(\mathbf{x}, t)$ is continuous on $\mathbb{X}$. (2) Show that $\mathbf{h}(\mathbf{x}, t)$ is continuous on $\mathbb{X}$ and t . (3) Show that $\mathbf{h}$ is one-to-one and onto.

Proof that $\mathbf{v}$ is continuous is done by contradiction. Suppose that $\mathbf{v}$ is not continuous. Then, from the PDE of (5.9) $\mathbf{b}$ would be unbounded, since the derivatives of noncontinuous functions are not continuous and unbounded. But from the definition of the body force in Section 4.2, the body force is bounded, so $\mathbf{v}$ must be continuous on $\mathbb{X}$.

Based on (5.2) and (5.4), $\mathbf{h}$ can be written as

$$\mathbf{h}(\mathbf{x}, t) = \mathbf{x} - \mathbf{x}(\mathbf{x}, 0) - \int_0^t \mathbf{v}(\mathbf{x}, \tau) d\tau \quad (5.14)$$

It is straightforward that $\mathbf{h}$ is continuous on $\mathbf{x}$ since it is a summation of functions that are continuous on $\mathbf{x}$. From Definition 5.1, $\mathbf{h}$ is continuous on t since $\mathbf{v}$ is bounded, say, by ϱ .

$$| \mathbf{h}(\mathbf{x}, t_2) - \mathbf{h}(\mathbf{x}, t_1) | = \left| \int_{t_1}^{t_2} \mathbf{v}(\mathbf{x}, \tau) d\tau \right| \quad (5.15)$$

$$< \int_{t_1}^{t_2} | \mathbf{v}(\mathbf{x}, \tau) | d\tau \quad (5.16)$$

$$< (t_2 - t_1) \varrho \quad (5.17)$$

From the existence theorem for ODEs [63], the solution to (5.4) exists and is unique. Additionally, for a fixed t , consider a complimentary velocity field $\tilde{\mathbf{v}}(\mathbf{x}, \tau) = -\mathbf{v}(\mathbf{x}, t - \tau)$ with solution $\tilde{\mathbf{u}}(\mathbf{x}, \tau)$ to the ODE

$$\frac{d\tilde{\mathbf{u}}(\mathbf{x}, \tau)}{d\tau} = \tilde{\mathbf{v}}(\mathbf{x}, \tau) \quad (5.18)$$

and boundary conditions $\tilde{\mathbf{v}}(\mathbf{x}) = 0$, $\forall \mathbf{x} \in \partial\mathbb{X}$ and $\tilde{\mathbf{h}}(\mathbf{y}, \tau) = \mathbf{y} - \tilde{\mathbf{u}}(\mathbf{y}, \tau) = \mathbf{y}$, $\forall \mathbf{y} \notin \mathbb{X}$. This solution also exists and is unique. Then, for all $\mathbf{x} \in \mathbb{X}$ we have $\mathbf{h}(\tilde{\mathbf{h}}(\mathbf{x}), t) = \mathbf{x}$ and $\tilde{\mathbf{h}}(\mathbf{h}(\mathbf{x}), \tau) = \mathbf{x}$. This implies that $\mathbf{h}(\mathbf{x}, t)$ is one-to-one with inverse $\mathbf{h}^{-1} = \tilde{\mathbf{h}}$.

Finally, suppose that for some t and $\mathbf{x} \in \mathbb{X}$, $\mathbf{h}(\mathbf{x}, t) = \mathbf{y} \notin \mathbb{X}$, then by definition of $\tilde{\mathbf{h}}$, $\tilde{\mathbf{h}}(\mathbf{y}, t) = \mathbf{y}$, but this cannot be since it has already been shown that $\tilde{\mathbf{h}}(\mathbf{y}, t) = \tilde{\mathbf{h}}(\mathbf{h}(\mathbf{x}, t), t) = \mathbf{x}$. Therefore, $\mathbf{h}$ must also be onto.

The algorithm for a discrete template deformation using the modified fluid model is the same as that of Section 5.3. In step 4., solution to the PDE of (5.9) is done using (5.13).

5.5 Fluid Model as a Prior Model

As mentioned at the introduction of this chapter, the fluid model attempts to maintain deformations which are meaningful or likely. This corresponds to the prior information about $\tilde{\mathbf{a}}$ which was neglected in (4.60). The model $\mathcal{L} = \mu_1 \nabla^2 - \mu_2$ is directly related to a prior probability through the use of the cost functional (5.8) and the Gibbs distribution (see Section 3.1.2). The functional corresponding to the differential operator of (5.9) can be related to a prior probability through the Gibbs distribution. Since $\tilde{\mathbf{a}} = \mathbf{a}(\mathbf{h}(\mathbf{x})) = \mathbf{a}(\mathbf{x} - \mathbf{u}(\mathbf{x}))$ it follows that

$$\begin{aligned} P(\tilde{\mathbf{a}}) &= P(\mathbf{h}) = P(\mathbf{u}) \\ &= \frac{1}{Z} \exp \left\{ -\frac{1}{2} \int_{\mathbf{x}} [\mu_1 |\nabla \mathbf{v}|^2 + \mu_2 |\mathbf{v}|^2] d\mathbf{x} \right\} \end{aligned} \quad (5.19)$$

$$= \frac{1}{Z} \exp \left\{ -\frac{1}{2} \int_{\mathbf{x}} [\mu_1 \left| \nabla_{\mathbf{x}} \frac{d\mathbf{u}}{dt} \right|^2 + \mu_2 \left| \frac{d\mathbf{u}}{dt} \right|^2] d\mathbf{x} \right\} \quad (5.20)$$

Analysis of (5.20) shows that for a discrete $\mathbf{u}$,

$$P(u_i | u_j, \forall j \neq i) = P(u_i | u_j, \forall j \text{ adjacent to } i) \quad (5.21)$$

Equation (5.21) is the definition of a Markov random field. Thus the fluid model is capable of capturing only the first-order dependencies between pixels. In fact, it is the case that almost all PDE-based deformable methods capture only the first- or second-order dependencies between pixels.

5.6 Implementation Issues

There are three constants associated with the template deformation: μ_1 , μ_2 , and Δ . From (5.11) and (5.7) we note that there is no loss of generality in choosing $\mu_1 = 1$, since any ratio of μ_2/μ_1 can be obtained by proper selection of μ_2 and the scale of $\mathbf{v}$ can be manipulated using the step size Δ . There is a trade-off in selecting Δ since for smaller step sizes, the approximation in (5.7) is more accurate but takes longer to deform. We use $\Delta = 1/(\max_{\mathbf{x}}(|\mathbf{v}(\mathbf{x})|))$. With a fixed μ_1 , μ_2 will stipulate the relative importance

between limiting the gradient of $\mathbf{v}$ and $\mathbf{v}$ itself. Lower values of μ_2 result in a smoother deformation, while higher values lead to a more fluid deformation. No attempt has been made to determine the optimum value of μ_2 , but values in the range 50-1000 are typically selected.

5.7 Experimental Validation

Figure 5.5 demonstrates that the original PDE introduces displacements orthogonal to the direction of the body force. In this figure a square was deformed to a rectangle. Figure 5.5(a) was deformed using the original PDE, and Figure 5.5(b) was deformed using the modified PDE.

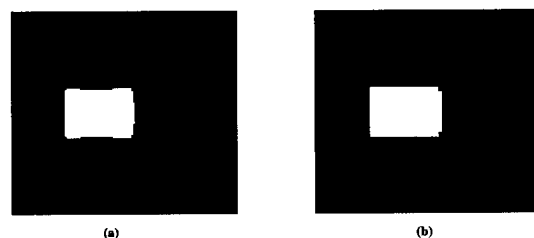


Figure 5.5 A third demonstration of orthogonally induced displacements. A square partially deformed to a rectangle using (a) the original fluid PDE, (b) the modified PDE. (a) shows a displacement in the top and bottom.

The final results for both the original and modified fluid models are nearly identical. Figure 5.6 shows a comparison of both methods.

Deformations resulting from the viscous fluid model are expected to preserve the topology of the template while allowing large, curved deformations. Figure 5.7 depicts the result of a patch to “C” [25, 66] deformation driven by a least-squares similarity metric. In this example, it is clear that removal of the divergence regularization for the velocity field from (5.8) and modification of the boundary conditions used to yield the explicit expression for the velocity field (5.13) have not removed the ability of the resulting deformations to accommodate large, curved displacements.

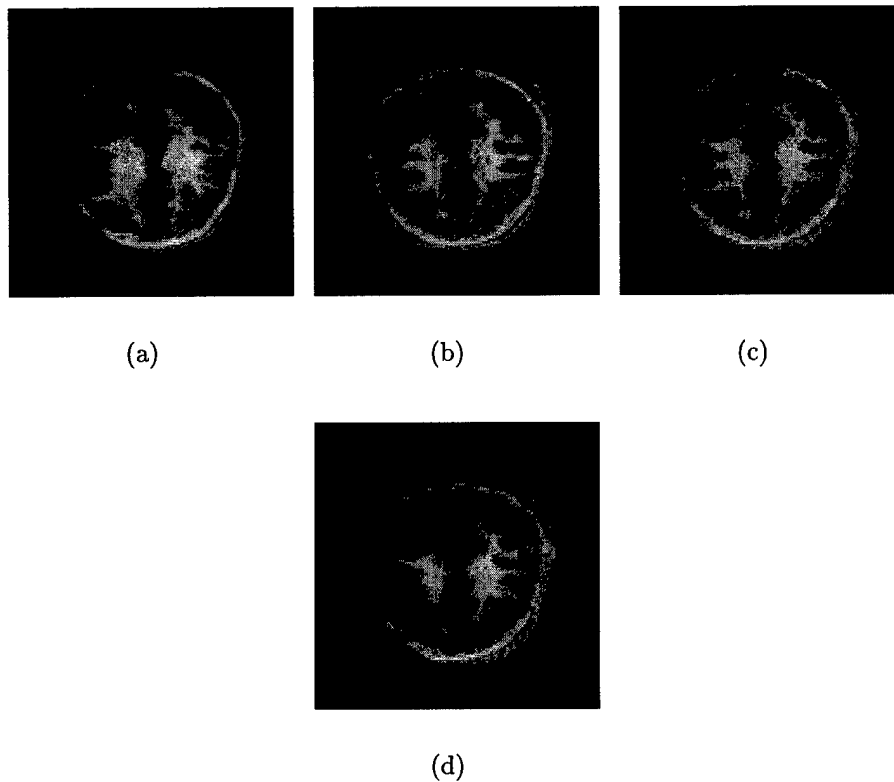


Figure 5.6 Comparison between the original and modified fluid models. (a) Template, (b) study, (c) deformed template using original PDE, (d) deformed template using modified PDE.

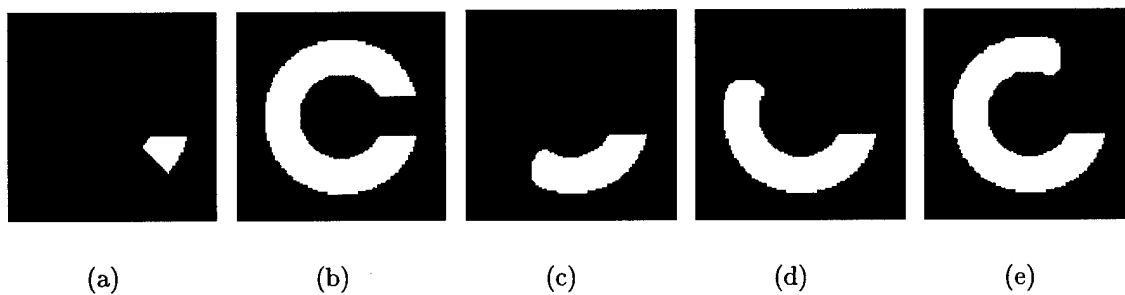


Figure 5.7 Illustration of the fluid transform's ability to accommodate large, curved deformations. (a) Template, (b) study, (c)-(e) deformed template after 50, 100, 150 time steps.

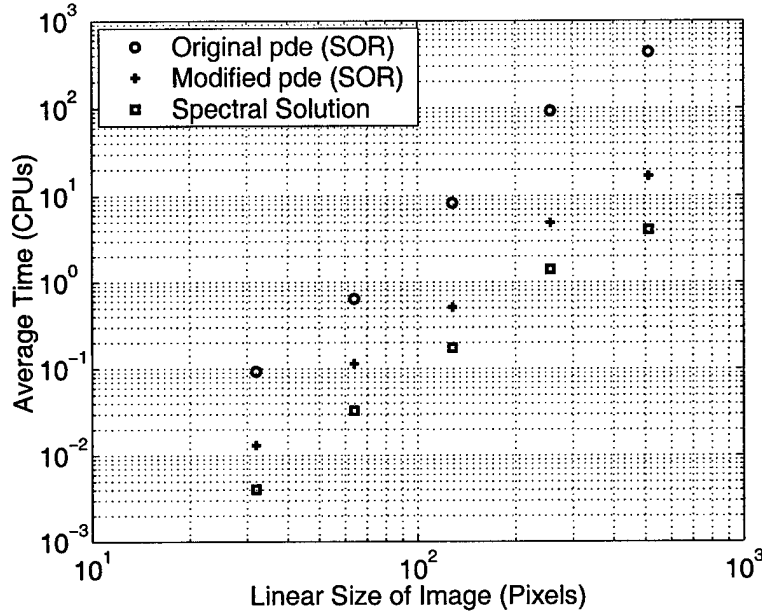


Figure 5.8 Average time required to solve for the instantaneous velocity field in the course of the patch to “C” experiment at various image sizes.

Attention is now turned to quantitative measures of the efficiency of the proposed technique. Figure 5.8 reports the average CPU time, in CPU seconds as measured on a 500-MHz ev56 DEC Alpha, for the numerical solution of (5.6) via SOR (“original pde”), the numerical solution of (5.9) via SOR (“modified pde”), and the spectral evaluation of the velocity field via (5.13) (“spectral solution”). In all cases, the time reported is averaged over the course of a patch to “C” experiment (Figure 5.7) performed at various image resolutions. For the PDE (5.6), it was found that the relaxation parameter $\omega \approx 1.9$ yielded optimal results. For the PDE (5.9), it was found that the relaxation parameter $\omega \approx 1.5$ yielded optimal results. The spectral solution method was implemented with fast transforms provided by the FFTW [67] package, which automatically determines the optimal FFT algorithm for a given data size and machine architecture.

It is clear from Figure 5.8 that the proposed spectral technique is indeed faster than the SOR method at solving either (5.6) or (5.9). Additionally, it can be seen that the greatest advantage is achieved by eliminating the penalty on the divergence of the velocity field. This accounts for the difference between the times indicated as “original pde” and

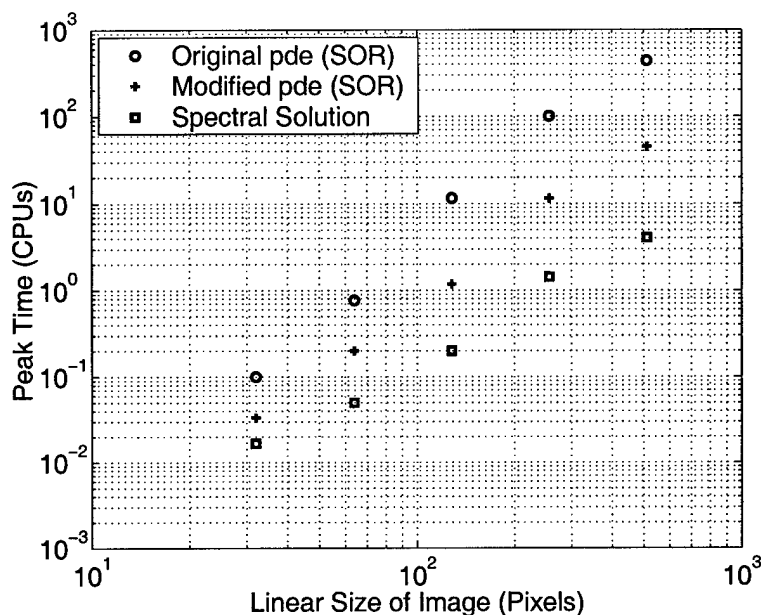


Figure 5.9 Peak time required to solve for the instantaneous velocity field in the course of the patch to “C” experiment at various image sizes.

“modified pde.” One surprising feature is that the spectral technique appears to scale no more favorably than the SOR solution of (5.9). This is explained by the fact that in the SOR procedure, the initial guess has a strong influence on the resulting number of iterations. In the case of modeling a time-variant velocity field, the field found at the previous discrete time step served as an excellent initial guess for the velocity at the current time step. This may not be the case in more complicated images.

Figure 5.9 reveals the expected superiority in the scaling of the spectral technique over the SOR by examining the maximum CPU times encountered in the course of the patch to “C” experiment. In these “worst-case” scenarios, the SOR loses the benefit of having a very good initial guess. All times are again in CPU seconds as measured on a 500-MHz ev56 DEC Alpha.

The performance of the modified technique has been examined on a 3-D example. Figure 5.10 depicts the result of a 3-D patch to “C” experiment. The experiment took place on a $128 \times 128 \times 32$ grid and was completed in under one hour on a 500-MHz DEC

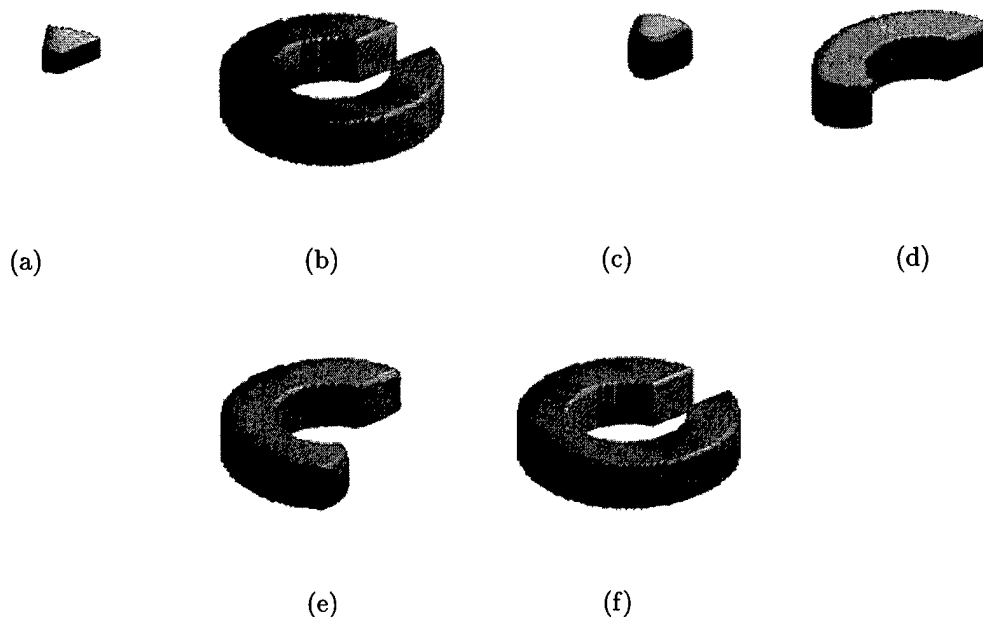


Figure 5.10 Three-dimensional patch to “C” experiment. (a) Template, (b) study, (c)-(f) template at various stages of the deformation.

Alpha ev5 CPU. Because the template is somewhat “thinner” than the target, truly 3-D deformations are induced.

Performance of the modified fluid model with mutual information has also been examined on medical images. Figure 5.11 shows a magnetic resonance (MR) proton density (PD) image from the Visible Human Project, a template and the resulting deformed template. The deformed template structure closely resembles the study image structure. Notice the ability of the deformation to track large, curved deformations.

Figure 5.12 shows a second demonstration of the deformation process for images with significantly different contrasts than Figure 5.11(a). In this case the top row shows deformation to an MR T-1 weighted image, and the second row shows deformation to a cryosection image. The same template and modified fluid deformation algorithm is used, and no preprocessing is necessary. Although Figure 5.12(a) is noticeably noisier than the other images, the algorithm still performs reasonably well.

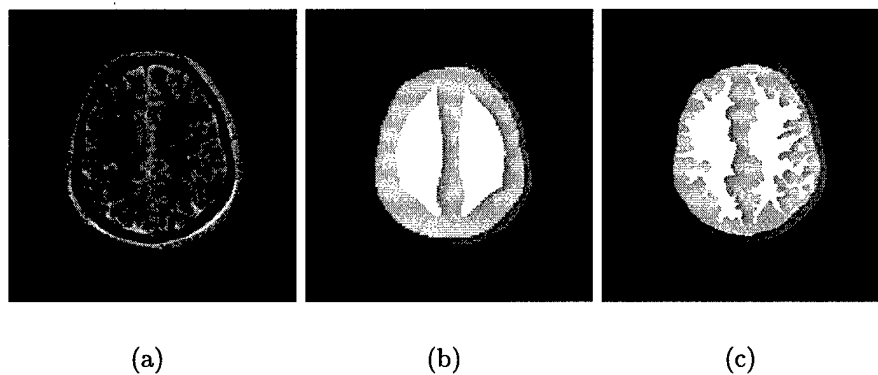


Figure 5.11 Template deformation to an MR-PD medical image. (a) Study, (b) template, (c) deformed template.

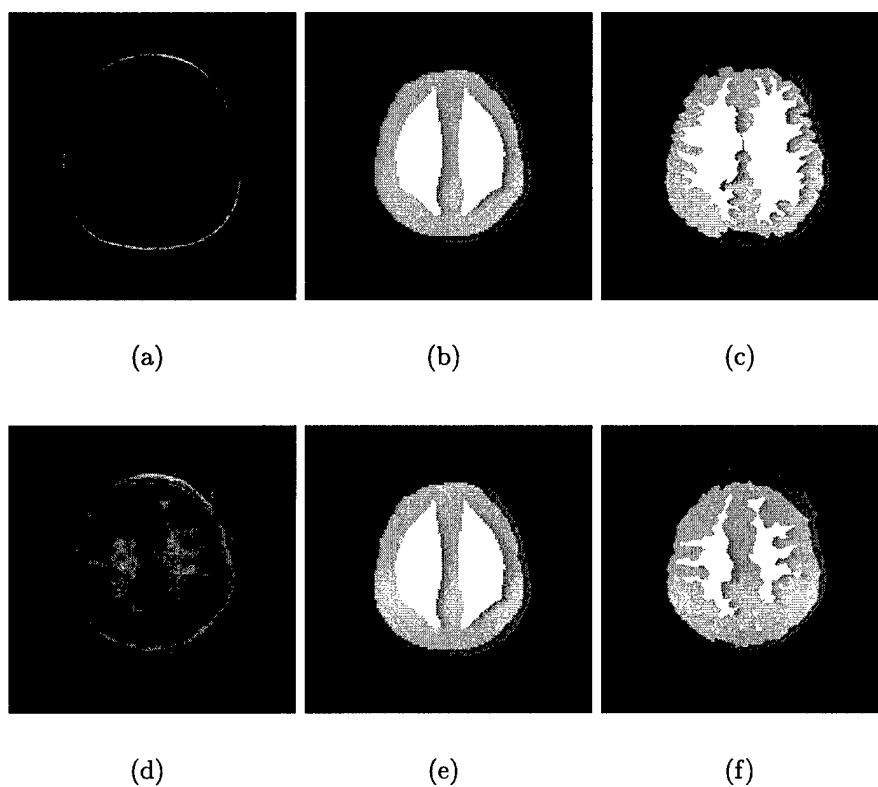


Figure 5.12 Deformation to MR-T1 and cryosection medical images. (a) Study, (b) template, (c) deformed template. (d) Study, (e) template, (f) deformed template.

Note that a large portion of the white matter (corresponding to the brightest portion of the template) at the lower left portion of the brain in Figure 5.12(a) is misclassified. This is partially due to initial alignment of the template, and also due to a topology difference between the template and study. For the deformation of the template to 5.12(d), the template intensity corresponding to skull (third darkest) over-deforms outward at the top of the image. This is a result of a slight gradient in the image intensity values from top to bottom. This is most likely due to lighting conditions at image acquisition.

Discussions earlier in this chapter and in Chapter 4 mentioned the theoretical impact of changes in the Parzen window width used in the body force calculation and changes in μ_2 from (5.9). Figures 5.13 through 5.15 show results for various settings of these parameters.

Section 4.1.3 stated that Parzen windowing should reduce partial volume and noise effects. In Figure 5.13 the first column shows the study images, the second column shows deformation using only normalized histograms, and deformations in the third and fourth columns use Gaussian Parzen windows with standard deviation 5 and 10, respectively. The results show that noise effects and partial volume effects are reduced. Figure 5.13(e) is quite noisy. The boundary between skull and scalp is quite jagged in (f) due to the noise, but these effects are reduced in (g) and (h). The fingerlike regions of the white matter typically suffer more from partial volume effects than other regions of the images. Comparison between columns of the fingerlike white matter regions shows that Parzen windowing allows the deformation to more fully extend the fingers in some cases.

Figures 5.14 and 5.15 show image deformations as μ_2 varies between 50, 100, and 1000 for columns 2, 3, and 4, respectively. Larger values of μ_2 lead to more fluid transforms. As the model becomes more fluid, the deformation is more capable of capturing the large nonlinear variations in the white/gray matter boundaries. However, other parts of the template are more likely to over-deform.

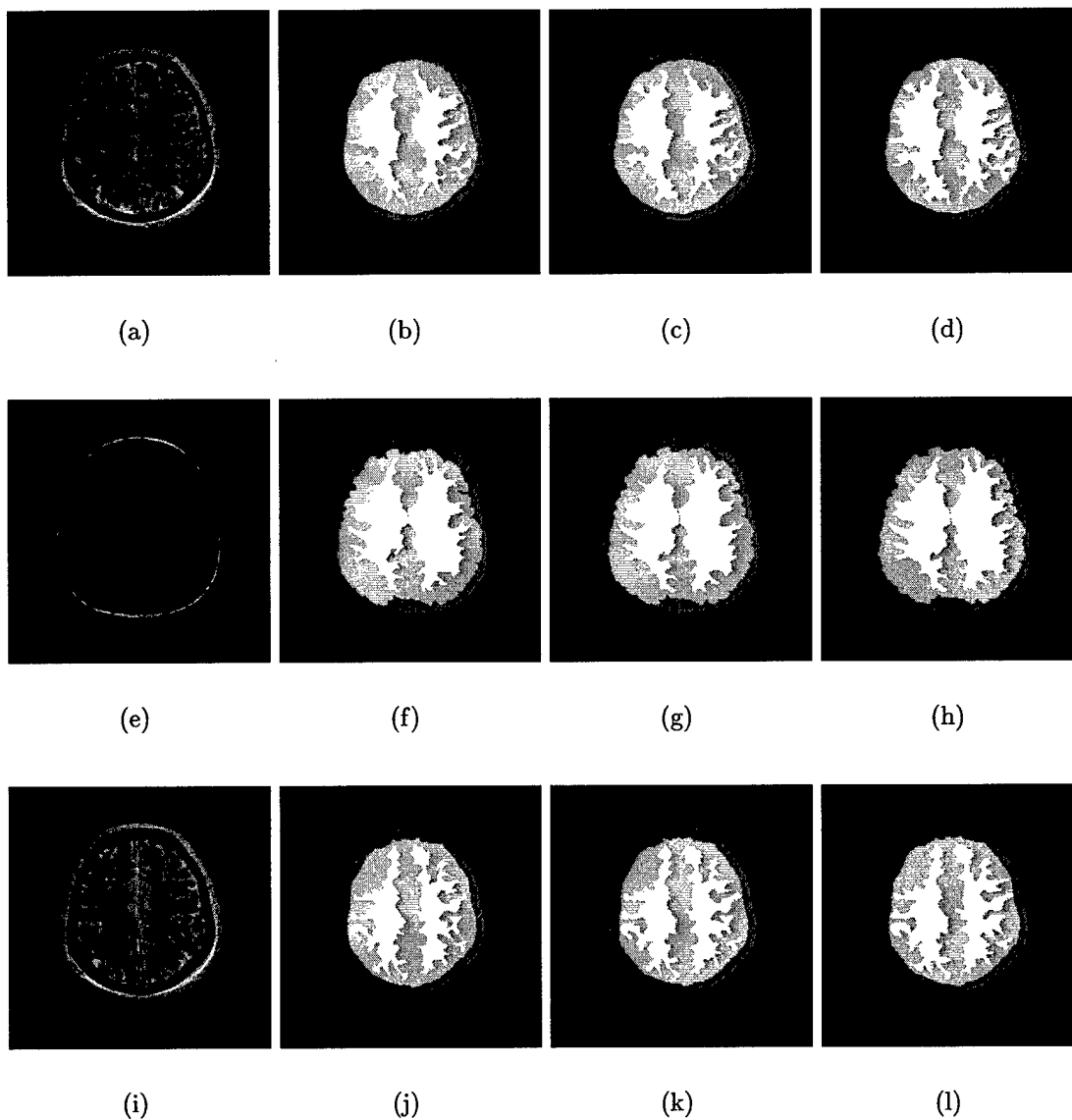


Figure 5.13 Deformation to several images using Parzen windows of differing widths when calculating the body force. From left to right, columns show the study, histogram-based probability estimation, Parzen estimation with $\sigma = 5$, and Parzen estimation with $\sigma = 10$.

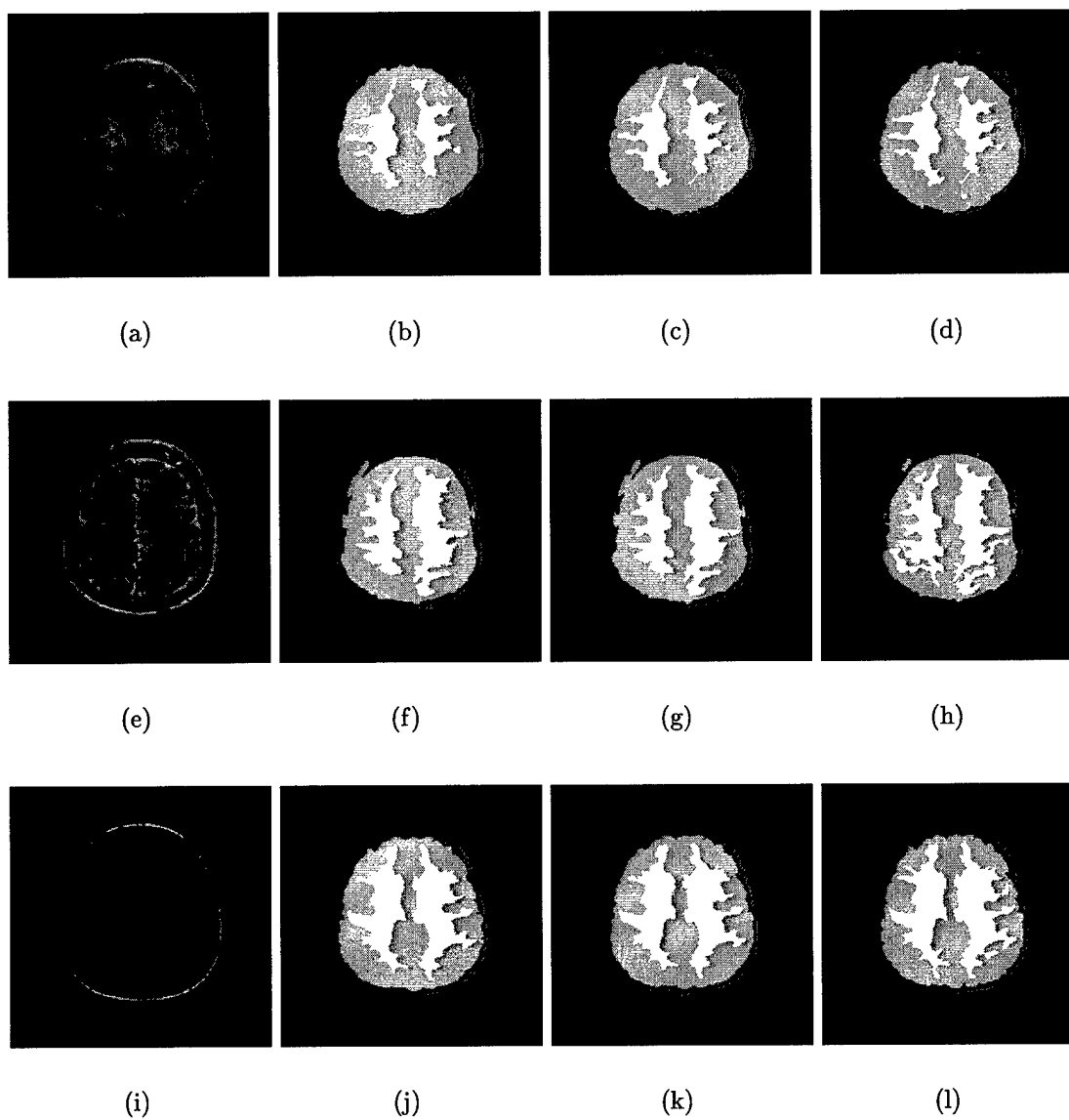


Figure 5.14 Deformation to several images using various values for μ_2 in (5.9). From left to right, columns show the study, and $\mu_2 = 50, 100, 1000$.

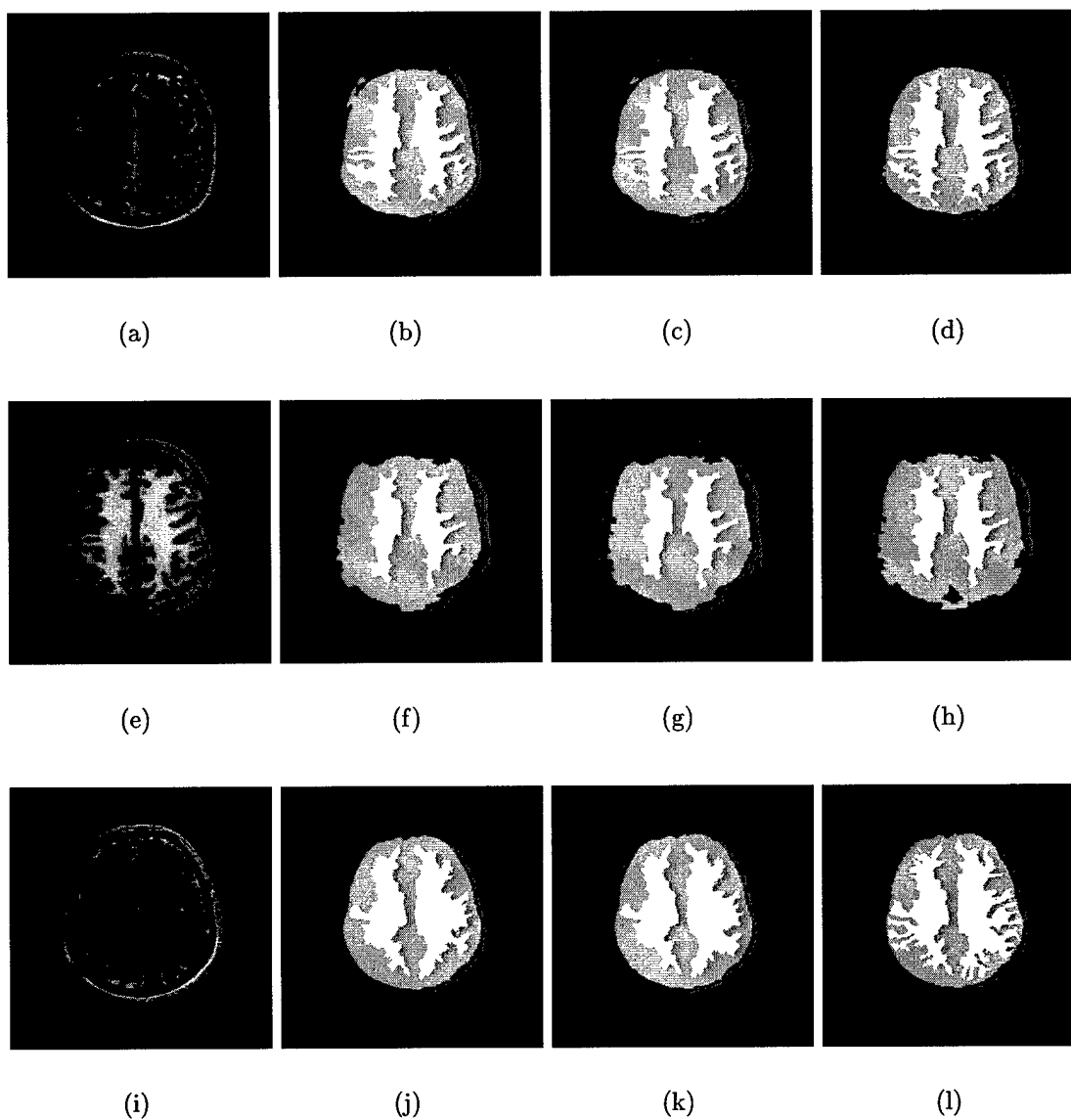


Figure 5.15 Deformation to several other images using various values of μ_2 . From left to right, columns show the study, and $\mu_2 = 50, 100, 1000$.

5.8 Summary

A fluid deformable model has been presented. The model maintains the topology of the template while allowing large, nonlinear deformations. The model is theoretically and experimentally faster than the fluid model proposed elsewhere. Parameters requiring manual selection were a fluid parameter μ_2 and the Parzen window width σ . A value of $\sigma = 5$ was found to perform well. When μ_2 was selected to make the deformations sufficiently fluid to capture the variability in the white/gray matter, the model tended to over-deform in other areas. The PDE imposes prior constraints about smoothness and topology on the template. The constraints were shown to be Markovian in nature. Use of more specific shape information is studied in Chapter 6.

CHAPTER 6

INCLUSION OF SHAPE INFORMATION

In allowing large curved deformations of the template, the fluid model also allows over-deformation of regions. A possible solution is to determine an allowable displacement for each element of the template; however, the displacements are not independent of each other. In fact, most interesting images, including brain images, contain dependencies between distant pixels. For example, the position of an element on one side of the skull may be coupled with a symmetric point on the opposite side of the image. These dependencies are critical to the generation of templates consistent with the objects they represent, but they are not well modeled using PDEs as developed in Chapter 5.

From an information-theoretic viewpoint, promoting likely variations corresponds to minimizing the information of the deformed template $I(\tilde{\mathbf{a}}(\mathbf{x}))$ (see Section 4.5.1). Section 5.5 showed that the fluid model is capable of modeling only the Markov dependencies in the template. This chapter discusses modeling dependencies which are higher order than Markov.

The shape modeling presented in this chapter is based on a point distribution model developed by Cootes et al. [40]. The point distribution model is summarized, followed by the presentation of a model for capturing global dependencies of a deformed template and a method for including the shape information in the deformation process.

6.1 The Point Distribution Model

Cootes et al. [39, 40] developed a model to capture meaningful, highly correlated variations among training samples and represent them in a compact, intuitive form: the point distribution model (PDM). The approach is to represent these dependencies in terms of an orthonormal basis set. Deformations are then restricted by simple independent limits on model parameters. The following is the PDM as described in [39, 40, 42].

A *shape* is defined to be an equivalence class for which instances of the class show a high degree of similarity through the operation of a similarity group (such as rotation, shift and scale). For purposes of the PDM, all shape instances must have the same number of points N_k and must be in the same order. For convenience, the point order usually follows sequentially around the shape, but this is not necessary. At times, a shape instance is also referred to as a shape.

Mathematically speaking, let $\mathbf{q} = \{q_k\}_{k=1}^{N_k}$ represent a shape instance. The similarity between $\mathbf{q}$ and an exemplar shape $\mathbf{q}^*$ is measured by [42]

$$\epsilon = \min_{\mathcal{T} \in \mathbb{T}} \sum_{k=1}^{N_k} (\mathcal{T}(q_{x,k}) - q_{x,k}^*)^2 + (\mathcal{T}(q_{y,k}) - q_{y,k}^*)^2 \quad (6.1)$$

where $\mathbb{T}$ is the set of affine transforms compensating for all possible translations, scales, and rotations of the shape denoted by $\mathcal{T}$.

Given a set of N_i shape instances, define $\mathcal{T}_i^*$ to be the $\mathcal{T}$ which minimizes ϵ for the i th shape instance. Each normalized shape instance $\mathcal{T}_i^*(\mathbf{q}_i)$ can be represented as a $2N_k$ -D vector. When all normalized training examples are plotted in the $2N_k$ -D space, they will form a cluster of points. This cluster lies in a region of space called the *allowable shape domain* (ASD). The PDM models the coordinate variations within the clusters. Assuming linear dependence on the coordinates, the ASD will be approximately ellipsoidal and the size will be approximately described by the cluster of training examples. Every vector within the ASD will correspond to a set of 2-D points whose shape is broadly similar to those in the training set.

The center of the ASD, or the average shape, is calculated using

$$\bar{\mathbf{q}} = \frac{1}{N_i} \sum_{i=1}^{N_i} \mathcal{T}_i^*(\mathbf{q}_i) \quad (6.2)$$

The ASD can be described using a basis set corresponding to its principal axes. Each axis gives a mode of variation, a way in which the 2-D shape points tend to move together as the overall shape varies. The principal axes are found using principal component analysis (PCA) [68]. Eigenvalues and eigenvectors are found for a covariance matrix

$$\Sigma = \frac{1}{N_i} \sum_{i=1}^{N_i} (\mathcal{T}_i^*(\mathbf{q}_i) - \bar{\mathbf{q}})(\mathcal{T}_i^*(\mathbf{q}_i) - \bar{\mathbf{q}})^T \quad (6.3)$$

The modes of variations within the training set are described by the unit eigenvectors $\mathbf{p}_k$ and the variance explained by each eigenvector is equal to the corresponding eigenvalue λ_k . Any point in the ASD ellipsoid can be reached by adding a linear combination of the eigenvectors to the mean shape. It turns out that almost all of the shape variations can be explained using the n_0 principal modes of variation; thus, any shape instance can be approximated using the mean shape and a linear combination of the first n_0 eigenvectors. Define

$$\Xi = (\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_{n_0}) \quad (6.4)$$

Then shape instance $\mathbf{q}$ is approximated by

$$\mathbf{q} \approx (\mathcal{T}^*)^{-1}(\bar{\mathbf{q}} + \Xi\boldsymbol{\beta}) \quad (6.5)$$

where $\boldsymbol{\beta}$ is a vector of weights. Due to the orthonormality of the eigenvectors, $\boldsymbol{\beta}$ can be found by

$$\boldsymbol{\beta}_i = \Xi^T(\mathcal{T}^*(\mathbf{q}) - \bar{\mathbf{q}}) \quad (6.6)$$

Suitable limits for $\boldsymbol{\beta}$ are typically on the order of

$$-3\sqrt{\lambda_k} \leq \beta_k \leq 3\sqrt{\lambda_k} \quad (6.7)$$

since most of a population lies within three standard deviations of the mean. The PDM, then, defines a shape using the mean shape instance and parameters Ξ and $\boldsymbol{\lambda}$.

6.2 The Shape Model

A single PDM is ill-suited to directly model all allowable variations in a template. This is due to two factors. First, large, nonlinear deformations of the template are allowed, but they are not well modeled using the PDM. Second, the dimensionality of the ASD for an entire template is cumbersome. Instead, PDMs are created for important subsets of the template. Dividing a complex shape into several, more well-behaved, subshapes is also problematic if correlation is ignored between points in different subshapes. For contour-based deformations, topology between subshapes is not necessarily conserved. A conflict could occur, for example, if image points belong to more than one subshape [42]. However, because the deformation of this thesis is volumetric rather than contour-based, multiple subshapes can be simultaneously deformed. The fluid properties of the deformation account for the topological constraints between subshapes.

Let $\mathbf{x}_j \subset \mathbb{X}$ be the j th of N_j subsets of pixel locations in the template. Each subset consists of $N_{j,k}$ points in $\mathbb{X}$. Through the deformation process these points are relocated to $\mathbf{q}_j = \mathbf{h}^{-1}(\mathbf{x}_j) \subset \mathbb{X}$. The overall shape which will be modeled is $\mathbf{q} = \bigcup_{\forall j} \mathbf{q}_j$. The subsets are selected such that expected variations are linear in nature, and each $\mathbf{q}_j$ has a high degree of regularity or shape as determined by ϵ in (6.1). However, variations in $\mathbf{q}$ may not be linear in nature.

Let $\mathcal{T}_j^*$ be the $\mathcal{T}_j$ which minimizes (6.1) for subshape $\mathbf{q}_j$. For a general 2-D transform allowing all possible shifts, rotations, and scales, each $\mathcal{T}_j^*$ can be described by four scalars and put in vector form $\mathbf{t}_j$. The transform vectors for each subset can be concatenated into a single vector $\mathbf{t}^* = t((\mathbf{t}_1^*)^T, (\mathbf{t}_2^*)^T, \dots, (\mathbf{t}_{N_j}^*)^T)^T$ of length $4N_j$. A rough approximation to the shape is obtained using only the mean subshapes and the rigid transforms.

$$\mathbf{q} \approx \bigcup_{\forall j} (\mathcal{T}_j^*)^{-1}(\bar{\mathbf{q}}_j) \quad (6.8)$$

The information content of the deformed template is roughly approximated by

$$I(\tilde{\mathbf{a}}(\mathbf{x})) \approx I(\mathbf{q}) = I(\mathbf{t}^*) \quad (6.9)$$

In addition to the affine variations in the subshapes, there are nonaffine variations which preclude ϵ in (6.1) from being zero. These variations are mathematically described by $\mathcal{T}_j(\mathbf{q}_j) - \mathbf{q}_j^*$. Ensuring that the generated deformations also follow likely nonaffine transformations leads to a more accurate approximation of $I(\tilde{\mathbf{a}}(\mathbf{x}))$ by

$$I(\tilde{\mathbf{a}}(\mathbf{x})) \approx I(\mathbf{t}^*) + \sum_{j=1}^{N_j} I(\mathcal{T}_j(\mathbf{q}_j) - \mathbf{q}_j^*) \quad (6.10)$$

6.2.1 Model statistics from training samples

Using the model of (6.10), the dimensionality of the probability space has been reduced from $2N_x N_y$ to $4N_j + \sum_{j=1}^{N-j} 2N_{j,k}$, where $N_{j,k}$ is the number of points in subshape j . The shape model can be made further amenable by the use of the PDM. The PDM is used to model both the intrashape variations (nonaffine variations within a subshape) and the intershape variations (correlations between affine transform parameters).

Consider a training set of N_i deformed templates $\{\tilde{\mathbf{a}}_i\}$, $i = 1, \dots, N_i$ from which $\mathbf{q}_{i,j}$ are obtained. Subscript i corresponds to the training sample, j to the subshape, and k to a particular point in the subshape. Let $\mathbf{t}_i^*$ be the parameter vector for the affine transforms which minimize (6.1) for all subshapes of the i th training example. Then define the mean of the $\mathbf{t}_i$ as

$$\bar{\mathbf{t}} = \frac{1}{N_i} \sum_{i=1}^{N_i} \mathbf{t}_i \quad (6.11)$$

For each transform, its deviation from the mean is

$$d\mathbf{t}_i = \mathbf{t}_i - \bar{\mathbf{t}}$$

and the covariance matrix is calculated for the relative position using

$$\Sigma_t = \frac{1}{N_i} \sum_{i=1}^{N_i} d\mathbf{t}_i d\mathbf{t}_i^T \quad (6.12)$$

Let

$$\Xi_t = (\mathbf{p}_{t,1}, \mathbf{p}_{t,2}, \dots, \mathbf{p}_{t,n_t}) \quad (6.13)$$

be the matrix containing the n_t principal eigenvectors of Σ_t . λ_t is the vector of corresponding eigenvalues.

Intrashape variations are captured in a manner similar to that described above for the intershape variations. For each of the N_j subshapes, a mean shape $\bar{\mathbf{q}}_j$, a principal eigenvector matrix Ξ_j , and a principal eigenvalue vector λ_j are found.

Any shape can be approximately described using the mean transform vector, the mean subshapes, and a weighted sum of the eigenvectors found in Ξ_t and Ξ_j , $j = 1, \dots, N_j$. The weights corresponding to the eigenvectors are β_t and β_j . When β_t and β_j are all zero vectors, the resulting shape is the average shape.

Once the model has been trained using the training data, the mean subshapes, eigenvalues, and eigenvectors are deterministic. The only uncertainty in the shape is found in the β_t and β_j . Thus, (6.10) becomes

$$I(\tilde{\mathbf{a}}(\mathbf{x})) \approx I(\beta_t) + \sum_{j=1}^{N_j} I(\beta_j) \quad (6.14)$$

Notation is slightly simplified by letting β_t and λ_t correspond to the $j = 0$ subshape. Equation (6.14) simplifies to

$$I(\tilde{\mathbf{a}}(\mathbf{x})) \approx \sum_{j=0}^{N_j} I(\beta_j) \quad (6.15)$$

The most likely templates are generated when (6.14) is minimized. This is done by minimizing the β parameters. Since the individual elements of β_j are assumed independent, they can be varied independently of each other. Acceptable ranges are determined by the λ_j

6.3 Inclusion of Shape Information into the Deformable Model

Under the proposed method, the learned shape information is introduced into the deformable process as an additional body force $\mathbf{b}_p$ as shown in (1.4), which is restated

here for convenience:

$$\mathcal{L}\mathbf{u} + \mathbf{b}_d + \mathbf{b}_p = 0 \quad (6.16)$$

where $\mathbf{b}_d$ is the body force from the mutual information, $\mathcal{L}$ is the fluid model, and $\mathbf{u}$ the displacement field.

Calculation of $\mathbf{b}_p$ is done in the following manner. Let $\mathbf{q}$ describe the current deformation at iteration time t . The associated $\beta_{j,k}$ give an indication of the prior information contained in the deformed template. Unless $\beta_{j,k} = 0$, $\forall j, k$, a more likely shape can be found by reducing some or all of the $\beta_{j,k}$. Denote a more likely shape by $\tilde{\mathbf{q}}$ with associated model parameters $\tilde{\beta}_{j,k}$. Following are some desirable properties of $\tilde{\beta}_{j,k}$:

1.

$$\tilde{\beta}_{j,k} \leq \beta_{j,k}, \quad \forall j, k \quad (6.17)$$

2.

$$|\tilde{\beta}_{j,k}| \leq 3\sqrt{\lambda_{j,k}}, \quad \forall j, k \quad (6.18)$$

3.

$$\left| \frac{\beta_{j,k}}{\sqrt{\lambda_{j,k}}} \right| \leq \left| \frac{\beta_{m,n}}{\sqrt{\lambda_{m,n}}} \right| \rightarrow \left| \frac{\tilde{\beta}_{j,k}}{\beta_{j,k}} \right| \leq \left| \frac{\tilde{\beta}_{m,n}}{\beta_{m,n}} \right| \quad (6.19)$$

4.

$$\text{sgn}(\tilde{\beta}_{j,k}) = \text{sgn}(\beta_{j,k}) \quad (6.20)$$

An equation for $\tilde{\beta}_{j,k}$ with these properties is

$$\tilde{\beta}_{j,k} = \text{sgn}\beta \cdot 3\sqrt{\lambda_{j,k}} \left(1 - \exp \left(\frac{-|\beta_{j,k}|}{\sqrt{\lambda_{j,k}}} \right) \right) \quad (6.21)$$

It follows that

$$\tilde{\mathbf{q}}_j = \tilde{\mathcal{T}}_j^{-1}(\tilde{\mathbf{q}}_j + \Xi_j \tilde{\beta}_j) \quad (6.22)$$

and the more likely shape is

$$\tilde{\mathbf{q}} = \bigcup_{j=1}^{N_j} \tilde{\mathbf{q}}_j \quad (6.23)$$

Finally, $\mathbf{b}_p$ is given by

$$\mathbf{b}_p = \mu_4(\tilde{\mathbf{q}} - \mathbf{q}) \quad (6.24)$$

6.4 Implementation Issues

The shape method just described introduces one constant, μ_4 for manual selection, in addition to those already explained in Section 5.6. The constant μ_4 represents the relative emphasis between the data-dependent body force and the shape-dependent body force. For all experiments shown in this thesis μ_4 was chosen such that the ratio between the maximum force in $\mathbf{b}_p$ and the maximum force in $\mathbf{b}_d$ was proportional to the maximum Mahalanobis distance of the subshapes (and transform vector) from their means. Specifically,

$$\mu_4 = \frac{\max_{\mathbf{x}}(|\mathbf{v}(\mathbf{x})|)}{\max_{k,j}(|\tilde{\mathbf{q}} - \mathbf{q}|)} \max_j \left(\sum_{k=1}^{N_k} \left(\frac{\beta_{j,k}^2}{\lambda_{j,k}} \right) \right) \quad (6.25)$$

The other parameters were chosen as explained in Section 5.6.

In order to implement the proposed model, the template subsets $\mathbf{x}_j$ to track as shapes must be determined. As mentioned previously, these subsets should have a high degree of regularity as measured by (6.1), and the variations should be linear in nature. In the current implementation, these points are manually selected. Future research could implement a more automated method for shape selection.

A second issue is that each training example $\mathbf{q}_i$ must have the same number of points k for all i , the points must be in the same order, and they must have the same start point. Manually segmenting a training image is already a time consuming task, and requiring the human expert to also count and order points along the segmentation surfaces magnifies the difficulty of the task. Duta et al. [69] present a method to solve this

problem which is fully automated and quite robust. The approach is based on a quasi-exhaustive exploration of point correspondences. An alternate method is chosen based on the deformation of the template $\mathbf{a}(\mathbf{x})$ to the *segmented structure* $\mathbf{s}_i(\mathbf{x})$ of the i th training image. Because the study images are of equivalent topology, a transform $\mathbf{h}_i(\mathbf{x})$ can be found such that $\mathbf{a}_i(\mathbf{h}_i(\mathbf{x})) = \mathbf{s}_i(\mathbf{x})$. Furthermore, $\mathbf{h}_i(\mathbf{x})$ is one-to-one, so each point in the deformed template can be mapped back to a unique point in the original template. Using these facts, an order is manually specified to the points comprising the shape in the original template. The template is deformed to all segmentations in the training set (without using shape priors), using $\mathbf{h}_i(\mathbf{x})$ to keep track of the order of the points as they deform.

The same procedure is used to keep track of the shape while the template is being deformed to a study image. In practice enough points of the template are given to describe the shape.

The MI-based body force $\mathbf{b}_d$ contains nonzero values on both sides of the region boundaries in the deformed template. It is important that either $\mathbf{b}_p$ is applied to both sides of the boundaries or that $\mathbf{b}_d$ is consolidated to one side of the boundary. Otherwise, $\mathbf{b}_p$ will not be able to sufficiently cancel elements of $\mathbf{b}_d$ which lead to over-deformation.

It is assumed that the template will remain aligned as to rotation throughout the deformation process, so the allowable transform space $\mathbb{T}$ for normalizing subshapes consists of three scalars specifying a scale and translation.

6.5 Experimental Results

Experiments were conducted to test the proposed method for segmentation of images with widely differing contrasts. The training set was comprised of 32 2-D axial slices of the supraventricular region of the brain. Images were obtained from the VISIBLE HUMAN project and through local sources. The images were manually labeled according to five regions: white matter, gray matter, bone, scalp, and background. A template was created and deformed to match each of the training image segmentations. The subshapes chosen

for the shape model were the outer contour of the scalp, the outer contour of the skull, and the inner contour of the skull. Both inter- and intrashape variations were estimated as described in Section 6.2.1. Enough principal components were used to capture 95% of the variance.

The template was then deformed to study images not found in the training set with the intent of estimating the correct segmentation of study images. Figures 6.1 and 6.2 show deformation results for images both with and without the inclusion of shape information. The figures illustrate the ability of the shape information to prevent over-deformation in areas around the brain while still allowing the large, curved deformations required to capture white/gray matter boundaries.

6.6 Summary

A shape model has been presented. Subsets of points in the template are chosen that show a high degree of regularity, or shape, when deformed. Both affine variations between subshapes and residual variations within each subshape are captured. Point distribution models are used to compactly model the dependencies between point and subshapes. By this method, the shape of the template is represented by an average shape instance and a set of independent parameters. Attenuation of these parameters leads to more likely shapes. By training the PDMs on a set of deformed templates, the number and ordering of points in the training set is automatically achieved.

The shape information is included in the deformation process through an additional body force component. The force drives the template toward more likely shapes based on the information learned in the shape model. An equation to calculate the body force was derived from an axiomatic characterization of the force.

Experiments were conducted to assess the ability of the shape information to maintain acceptable transformations of the template. The inclusion of shape information was qualitatively shown to maintain acceptable deformations of the scalp and skull contours, while allowing large, nonlinear deformations of the white matter and gray matter.

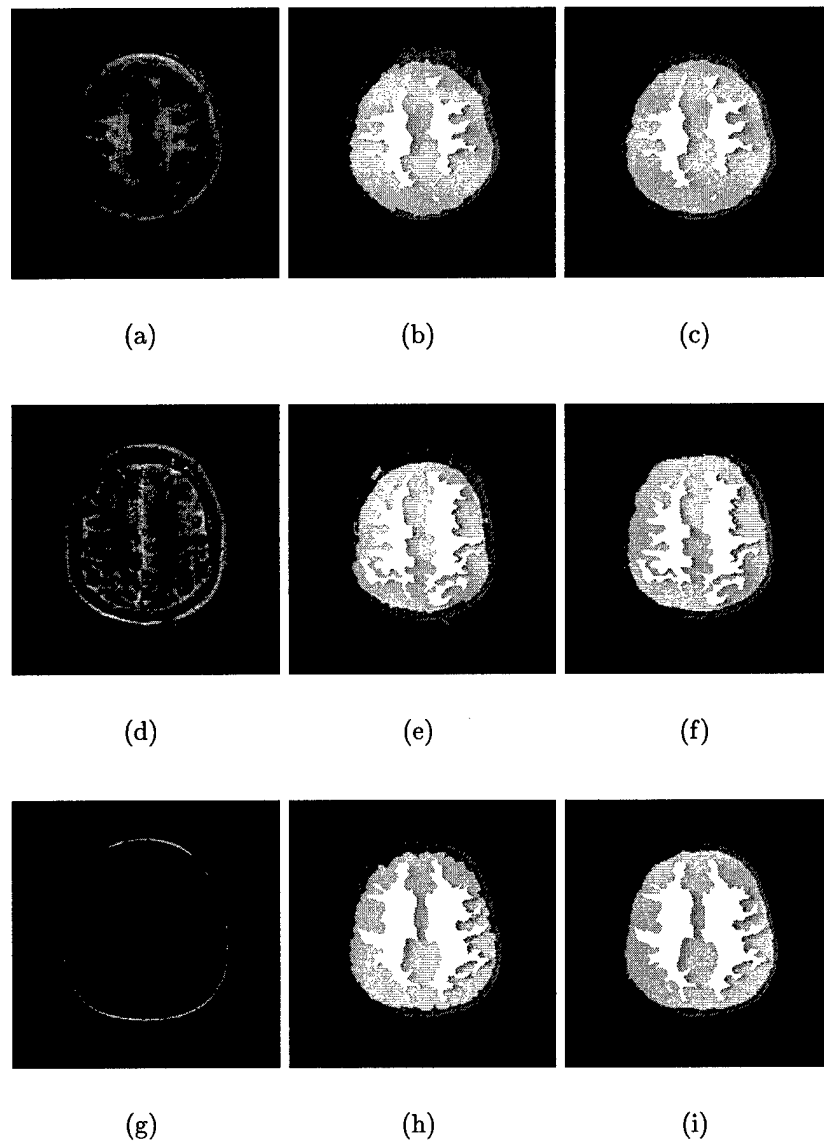


Figure 6.1 Demonstration of the effectiveness of shape information in the deformation process. Column 1: Study images. Column 2: Deformation without shape information. Column 3: Deformation with shape information. The shape information maintains better control of scalp and skull contours.

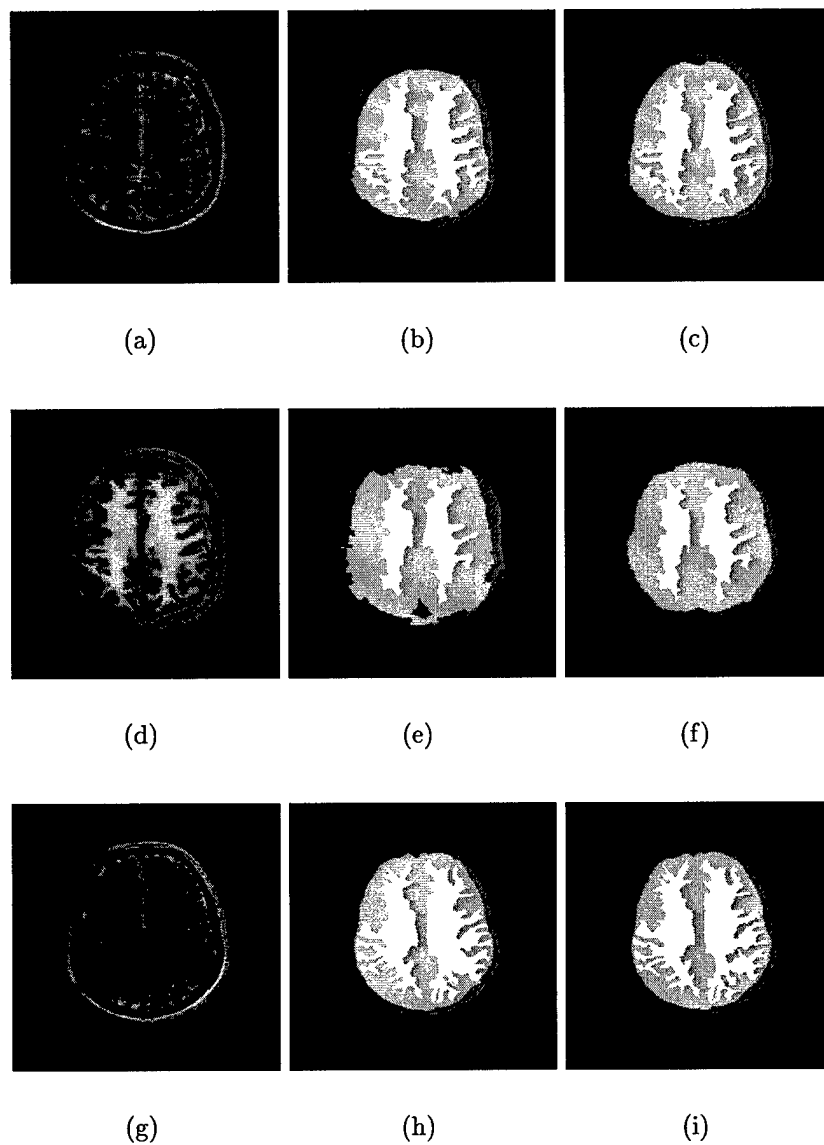


Figure 6.2 Demonstration of the effectiveness of shape information in the deformation process (continued). Column 1: Study images. Column 2: Deformation without shape information. Column 3: Deformation with shape information.

CHAPTER 7

A HYBRID BRAIN IMAGE SEGMENTATION METHOD

The deformable template method described in Chapters 4, 5, and 6 has been used as part of an automatic brain image segmentation framework being developed at the University of Illinois Beckman Institute [6]. This chapter briefly introduces the framework, shows how template deformation fits in, and shows experimental results for white/gray matter segmentation in MR images.

7.1 Brain Image Segmentation

Over the last decade, many methods have been proposed to tackle this problem [2, 59] including low-level, knowledge-based, and statistical techniques. Some of the earliest image segmentation techniques involved primarily low-level computer vision techniques such as edge detection [59, 70, 71]. Edge detectors are often not too useful for brain image segmentation because they identify local edges based on image intensity gradient information alone, and as a result, detected edges are usually not well connected to form anatomically meaningful regions. Region-based segmentation methods [72, 73, 74, 75, 76, 77] overcome some of these problems, but still do not usually contain sufficient prior knowledge.

Knowledge-based segmentation techniques, on the other hand, rely primarily on prior knowledge of the anatomical structures. Examples include model-based methods [78, 79]

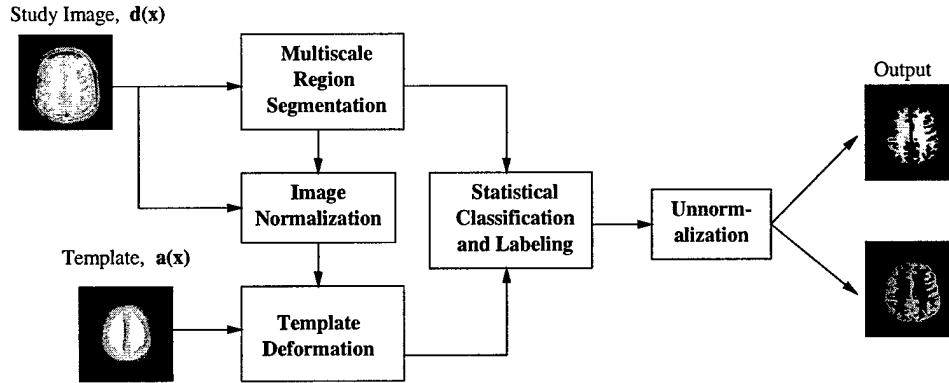


Figure 7.1 Flowchart of the described brain image segmentation method.

and rule-based methods [80]. Model-based methods fit a specific geometric model of the brain to a given data set to determine the final segmentation. This approach works well for some normal brain images but has trouble dealing with brain abnormalities. Rule-based methods use “if-then” type heuristic rules for pixel clustering and their practical utility is often limited due to inflexibility.

Other methods attempt to incorporate *a priori* information in a flexible manner using a statistical model [81, 82, 83] or a neural network model [84]. However, many computational problems remain, which limit their practical application.

The segmentation method described here takes advantage of several existing methods; the framework employs multiscale region segmentation, image normalization, template deformation, and statistical classification for the segmentation task. The resulting method can be viewed as performing region clustering based on geometric considerations. Figure 7.1 schematically shows the overall process of the proposed method. The template deformation component is that described in Chapters 4, 5, and 6. The other components are described below.

7.1.1 Feature extraction: Multiscale region segmentation

Due to the nature of MR imaging, different brain tissue types tend to have different intensities in the images. Intensities cannot be used exclusively in segmentation because of field inhomogeneities, partial volume effects, and noise. However, intensity information

tends to be a very useful feature in the segmentation process. The purpose of multiscale region segmentation is to organize the intensity information by grouping image pixels according to a similarity measure based both on their spatial relationship and intensity variations at different scales.

More formally, a homogeneous region $R(j)$, $j = 0, 1, 2, \dots$ with homogeneity scale σ is a set of connected pixels with the following properties:

1. $\|d(x_1), d(x_2)\| < \sigma, \quad \forall x_1, x_2 \in R(j);$
2. $\|d(x_1), d(x_2)\| > \sigma, \quad \forall x_1 \in R(i), \forall x_2 \in R(j), i \neq j$

where $\|\cdot, \cdot\|$ is some homogeneity measure which accounts for both intensity differences and spatial distances between points. One possible measure is the absolute intensity difference between connected pixels, given by

$$\|d(x_1), d(x_2)\| = |d(x_1) - d(x_2)|$$

Practically, however, applying such a definition directly will not usually give meaningful regions due to noise and other sources of image intensity variations [74, 75]. Most existing methods attempt to reduce the effect of the noise by image smoothing techniques. An alternate approach was proposed by Ahuja [76] and Tabb and Ahuja [77]. This method clusters pixels using attraction force fields, based on both spatial and intensity relationships with other pixels in the image, and explicitly provides for multiple scale clustering, based on the many levels of sensitivity. A main advantage of this algorithm is that it does not involve smoothing of the image to generate different scales, so boundary detail is maintained, even in the coarse scales. Moreover, gradients are not used, so noise is not amplified. The resulting segmentations are organized into a parent-child tree structure where each region in a coarser level consists of several subregions in the finer level. The tree structure allows for faster analysis due to the linking of information between scale levels.

To perform the multiscale analysis on the study image $\mathbf{d}$ a vector field $F(\mathbf{x}, \sigma_g, \sigma_s)$ is created for each desired scale by

$$F(\mathbf{x}, \sigma_g, \sigma_s) = \sum_{\hat{\mathbf{x}} \in \mathcal{X}} d_g(\Delta G, \sigma_g) d_s(\hat{\mathbf{x}} - \mathbf{x}, \sigma_s) \frac{\hat{\mathbf{x}} - \mathbf{x}}{\|\hat{\mathbf{x}} - \mathbf{x}\|}.$$

Here σ_g and σ_s are the scale parameters for graylevel and spatial similarities, $\mathbf{x}$ and $\hat{\mathbf{x}}$ are element locations in $\mathbf{d}(\mathbf{x})$, ΔG is the difference in intensities between points $\mathbf{x}$ and $\hat{\mathbf{x}}$, d_g is the intensity distance measure, and d_s is the spatial distance measure. The resulting vector field contains a vector for each image location that points toward other image elements to which it is 'close' in both intensity and position.

The intensity distance measure d_g is computed using a boxcar function such that

$$d_g(\Delta G, \sigma_g) = \begin{cases} 1 & \Delta G \leq \sigma_g \\ 0 & \text{otherwise.} \end{cases} \quad (7.1)$$

The spatial distance measure d_s is also a boxcar function similar to (7.1).

Tabb and Ahuja [77] state that having two scale parameters (σ_g and σ_s) is overly general. A meaningful value for σ_s can be adaptively calculated based on the value of σ_g and the image content. The goal is to select a small enough region of interest so that important image information is not lost while still providing meaningful region information. Exact details are provided in [77]. Using this adaptive procedure, $F(\mathbf{x}, \sigma_g, \sigma_s)$ is reduced to $F(\mathbf{x}, \sigma)$ where σ is the original σ_g .

Region information is extracted from $F(\mathbf{x}, \sigma)$ by realizing that the vectors within the field will tend to be perpendicular to region boundaries and converge toward the centers of the regions. Region boundaries are located between diverging vectors in the vector field. The regions for the different scales are organized in a tree structure so that regions at a coarse scale are composed of one or more regions in the next finer scale. Figure 7.2 shows the region segmentation for three different scales.

The region segmentation is also useful in automatically separating the head image from the background information. An object silhouette is determined by selecting a coarse scale and performing morphological fill operations followed by a closing operation.

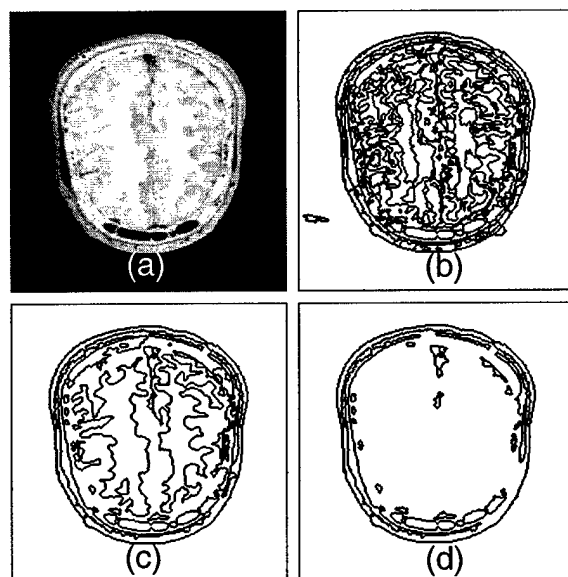


Figure 7.2 Multiscale analysis performed on an image. (a) Image, (b) $\sigma = 20$, (c) $\sigma = 30$, (d) $\sigma = 40$.

The region information and fill operations isolate the object from the background, and the closing operation eliminates any small holes in the silhouette. The silhouette is useful for image normalization, as will be discussed in the next section.

7.1.2 Geometric constraints

Although the multiscale region analysis provides useful features for brain image segmentation, these features alone are inadequate for forming a meaningful anatomical brain image segmentation. For example, without prior knowledge of the object being imaged, it is unclear which scale levels contain the proper region information and how these regions should be combined or split to create a meaningful anatomical structure. Often used techniques use region growing, split-and-merge, and linking methods [59, 75, 80, 85] to form a segmentation from regions. These methods typically use pixel intensities, gradients, and region similarity as the criteria for combining regions, so the final segmentation may be geometrically incompatible with the prior knowledge about the system.

In contrast, the method described herein relies on geometric information to constrain the segmentation process. In our case the geometric prior information is in the form of a

template and a deformable model. The first step in applying the template to the image is image normalization.

Image normalization

Normalization is achieved by a rigid-body (affine) registration of the template and study image. The practice of using an initial low-dimensional transformation (such as affine) followed by a more localized transformation is widely used [24, 26, 35, 86] due mainly to the fact that it overcomes local minima and leads to faster convergence.

Many rigid-body registration techniques, including those compared in [87], are iterative techniques that adjust for only rotation and shift differences between the two images. In [24], however, Bajcsy and Kovacic use a noniterative method which compensates for shift, rotation, and scale. This method is usually not as accurate as the iterative methods, but since normalization will be followed by non-rigid-body deformation, speed and simplicity are the main concern here.

The algorithm calculates a normalizing affine operator $\mathcal{N}_d$ which transforms $\mathbf{d}$ to a normalized size and orientation. The calculation consists of several steps. First, a binary silhouette image $\mathbf{d}_b(\mathbf{x})$ is created from the multiscale region information (see Sections 7.1.1 and 7.3). Second, the mass center of the silhouette is moved to the image origin to remove the translational offset. Next, major and minor axes are found using distances of silhouette pixels from the origin. The axes are calculated using an eigenvector decomposition of a covariance matrix for the silhouette locations. Finally, a rotation is calculated that will align the image to a normalized orientation. These steps are now formalized in detail.

1. Move the mass center of the silhouette $\mathbf{c} = (c_x, c_y)$ to the origin, where

$$c_x = \sum_{\mathbf{x}} x \mathbf{d}_b(\mathbf{x})$$

$$c_y = \sum_{\mathbf{x}} y \mathbf{d}_b(\mathbf{x})$$

2. Calculate the eigenvalues, namely λ_1, λ_2 , and eigenvectors (with unit length), namely v_1, v_2 of a covariance matrix M where

$$M = \begin{bmatrix} u(2,0) & u(1,1) \\ u(1,1) & u(0,2) \end{bmatrix}$$

and $u(k, r)$ is the central moment defined as

$$u(k, r) = \sum_{\mathbf{x}} (x - c_x)^k (y - c_y)^r \mathbf{d}_b(\mathbf{x}) \quad (7.2)$$

3. Rotate the image according to the eigenvectors of M so that covariance matrix is diagonalized:

$$\mathbf{x}' = A(\mathbf{x} - \mathbf{c})$$

where

$$A = [v_1 \ v_2]^T \quad \text{or} \quad A = [v_1 \ -v_2]^T$$

such that the diagonal elements have the same sign.

With the above transform, $\mathbf{d}(\mathbf{x}')$ becomes invariant to translation, shift, and scaling and is in rough alignment with the template; template deformation as described in this thesis can be performed. The deformation process will bring the template into better geometrical alignment with the structure of the study image.

7.1.3 Final classification

The segmentation given by the deformed template may not be accurate for several reasons, including the following: (1) It may converge to a local minimum; (2) finding the global minimum may be impractical; (3) even if the globally optimal solution is obtained, there is no guarantee that the estimate is optimal with respect to natural measures such as the human eye; and (4) topological differences may preclude correct segmentation. The purpose of the final classification is to combine information from the multiscale region

analysis with that obtained through template deformation to obtain a more meaningful result.

Final classification is achieved by defining a statistical model which incorporates information from both the deformed template and the multiscale pyramid. Each pixel will then be classified according to this model. The probabilistic model, denoted as $P(i, k)$, is the confidence that the i th image pixel belongs to class k . First, let $P^{(j)}(i, k)$ define the confidence that the i th pixel belongs to k based on the information in scale j .

$$P^{(j)}(i, k) = \frac{\text{Area}(R^{(j)}(i) \cap S_k)}{\text{Area}(R^{(j)}(i))} \quad (7.3)$$

where $R^{(j)}(i)$ represents the region in the j th level which contains pixel i , and $S_k = \{i \mid \tilde{a}_i = k\}$. $P(i, k)$ is the weighted sum of all scales where the weights $\omega_j \geq 0, j = 1, 2, \dots, J$, convey the confidence placed in each scale level. Then

$$P(i, k) = \sum_{j=1}^J \omega_j P^{(j)}(i, k). \quad (7.4)$$

If ω_j are restricted such that $\sum_{j=1}^J \omega_j = 1$, then

$$\sum_{k=1}^K P^{(j)}(i, k) = 1 \quad (7.5)$$

and

$$\sum_{k=1}^K P(m, k) = 1 \quad (7.6)$$

so both $P^{(j)}(i, k)$ and $P(i, k)$ are valid probability mass functions.

Once $P(i, k)$ has been calculated, the final segmentation $\hat{s}$ of the study image $\mathbf{d}$ is written as

$$\hat{s}(i) = \arg \max_k P(i, k) \quad (7.7)$$

Notice that the $P^{(j)}(i, k)$ maintain the region information of both the multiscale analysis

and the deformed template; for example,

$$P^{(j)}(i, k) = \begin{cases} 0 & \text{if } R^{(j)}(i) \cap S_k = \emptyset \\ 1 & \text{if } R^{(j)}(i) \subset S_k \end{cases} \quad (7.8)$$

7.2 Implementation Issues

The final classification was found to be sensitive to values of the homogeneity scale parameter σ used in the multiscale region analysis. Using many small values of σ (e.g., 1, 2, 3, 4, 5, 6, 7) and few large values of σ (e.g., 30, 40) can cause an improper bias by relying too much on the finer scales. Furthermore, values of sigma which are appropriate for a particular image contrast may not be useful for a different input image with a different contrast. In general, values for σ should be selected so that significant structural differences exist for different values of σ . One such measure is to monitor the number of distinct values the region grayscale averages take on. Using this measure, values are chosen such that the number of distinct region averages strictly decreases for larger values of σ used. In [77], Tabb and Ahuja discuss structures that are perceptually relevant based on the stability of the structures when σ is changed. Similar analysis could be done to decide which σ values should be used for our purposes. This is an area for further research.

Once useful values for σ have been found, the confidence placed in scale level, as denoted by ω_j in (7.6), is determined. Assuming that the deformed template result is a fairly good estimate of the segmentation, ω_j is set to be proportional to the inverse of the average number of S_k spanned by the deformation regions in level j . Thus, more confidence is placed in scales that span fewer regions. This has the effect of weighting the finer scales slightly more heavily than the coarse scales. Results for other values of ω_j , such as uniform weighting, have not been found to dramatically change the classification result. Other machine learning algorithms are being investigated to determine the suitability of each multiscale level. One such method is the use of neural network training of the values of ω_j .

7.3 Experimental Results

Experiments were conducted using optical cryosection and magnetic resonance (MR) brain images from the *VISIBLE HUMAN* database and MR brain images from [88] and local sources.

A test was conducted to assess the ability to extract the outer contour of the head from the image using the multiscale region analysis. The outer contours of 610 2-D 256×256 cryosection and MR images were estimated. Each image was segmented using multiscale region analysis. The scale corresponding to $\sigma = 30$ was operated on, using morphological fill operations and a closing operation to separate the head from the background and remove small gaps in the contour. One set of images from a local source, comprising about 20% of images, was found to not fully utilize the dynamic range of the grayscale space $\{0, 255\}$, so the set was linearly scaled to use the entire range. The other images were not preprocessed.

Since the contour is used for image normalization purposes, the silhouette extraction performance was measured by assessing the deviation of the normalized image from the 'correct' orientation as determined by a human observer. Contours for which the normalization procedure was able to transform the image to within 3° of rotation, or 3% of the image size in translation, shift, or scale of the correct orientation were considered correctly normalized. Of the 610 images, the contour feature was sufficiently extracted to correctly normalize all but 10 of the images. The processing was modified to allow $\sigma = 20, 30$, or 40 , followed by manual selection of the best one. Using this method, all but two of the images were correctly normalized. Figure 7.3 shows an image and an incorrect silhouette using $\sigma = 30$ but a correct silhouette using $\sigma = 20$. In this case selecting too coarse a scale causes part of the head to be considered background. The final two images required manual intervention to be correctly normalized.

Tests were conducted on the ability of the segmentation method to extract white matter and gray matter regions from a brain image. The template was deformed to each of the images using the deformable method outlined in this thesis (deformable result).

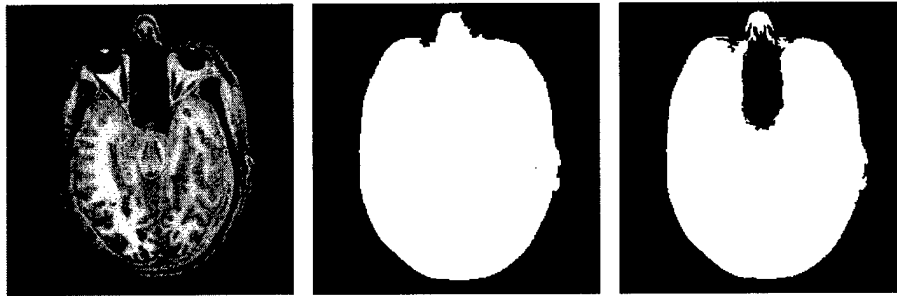


Figure 7.3 Extracting the brain silhouette. Left: An image, Center: the silhouette extracted using $\sigma = 20$, and Right: $\sigma = 30$. $\sigma = 20$ is the better choice for this image.

The deformable result was combined with a multiscale region segmentation result as outlined in Section 7.1.3 (combined result). Figures 7.4 and 7.5 show results from experiments conducted. The first column in each figure shows the study images, the second column shows the template deformation result, and the third column shows the final classification based on the merging of both the template deformation and multiscale analysis. In Figure 7.5, the final column shows the correct manual classification as done by an expert [88]. As is evident from the results, the combined result improves the segmentation in some areas, but seems to degrade other areas. The combined result improves the deformation in some of the fingerlike regions, where the deformation is not able to entirely deform. In the images of Figure 7.5, the combined result overextends the white matter portions.

Areas of the manual segmentations (column 4 of 7.5) show deep grooves in outer portion of the gray matter. These areas correspond to cerebral spinal fluid (CSF). The template contains five graylevels to capture scalp, skull, gray matter, white matter, and background. In reality there are many more tissue types present, including CSF, fat, muscle, and blood. As explained in Section 4.5, using too many graylevels in the process leads to deformation instability. Thus, there is a trade-off between stability and the number of tissues tracked.

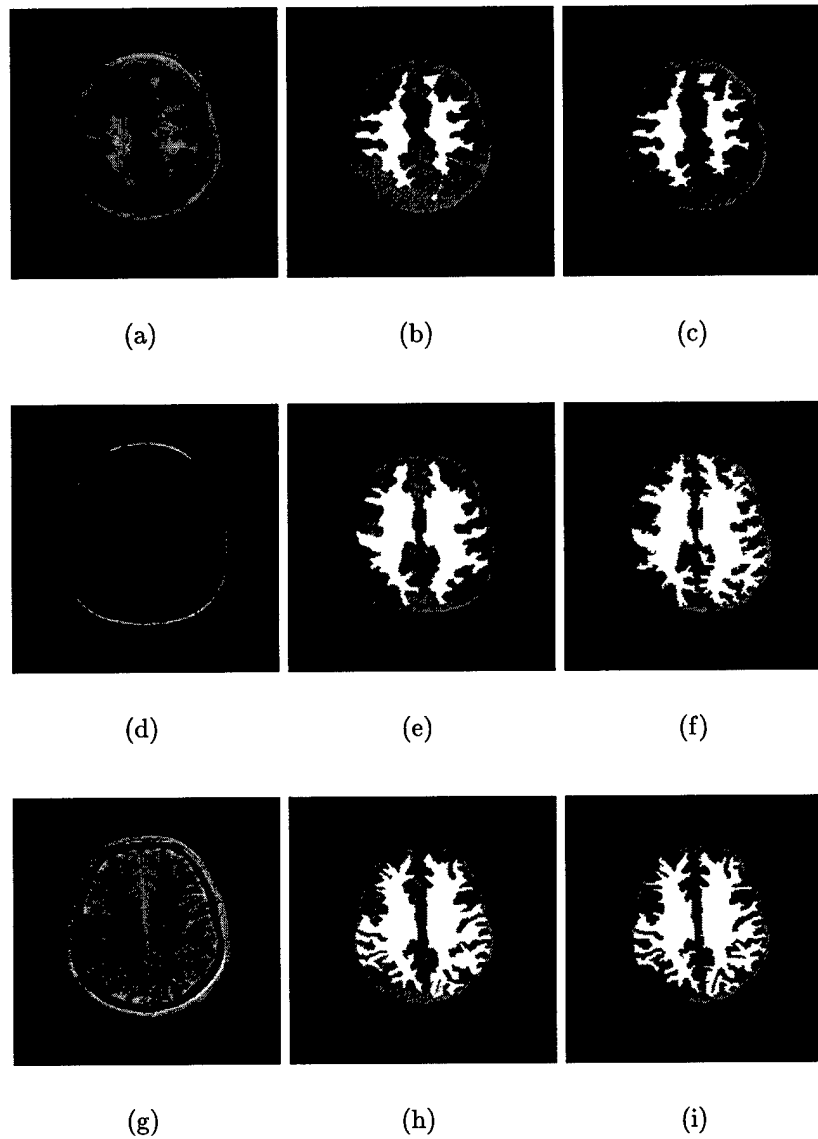


Figure 7.4 Comparison between segmentation based on template deformation and the combination of template deformation and multiscale region analysis. Column 1: Study images. Column 2: Template deformation. Column 3: Segmentation combining deformation and region analysis.

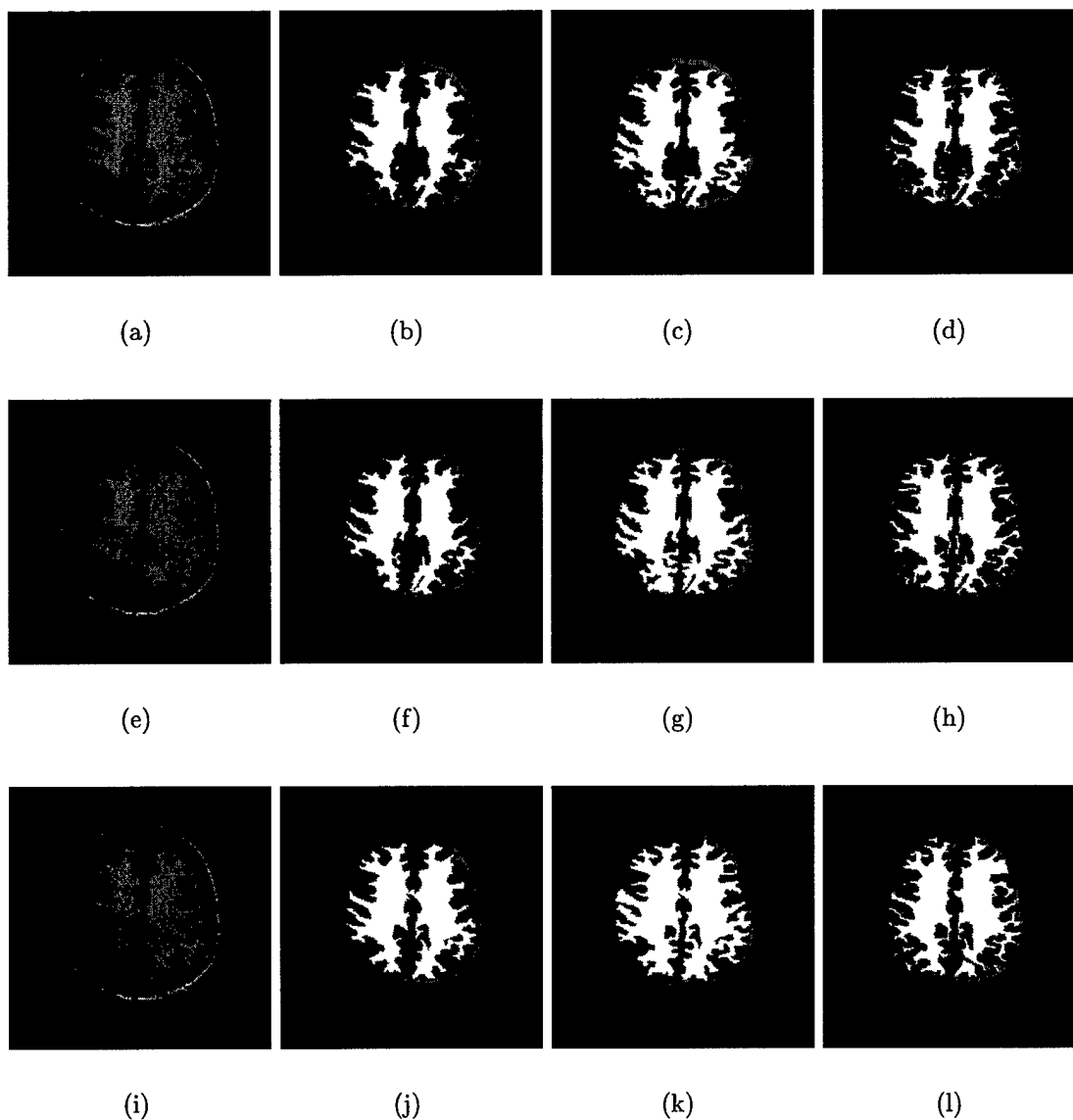


Figure 7.5 Comparison between segmentation based on template deformation and the combination of template deformation and multiscale region analysis (continued). Column 1: Study images. Column 2: Template deformation. Column 3: Segmentation combining deformation and region analysis. Column 4: Manual segmentation result.

7.4 Summary

This chapter presented a segmentation framework which is under development at the University of Illinois Beckman Institute. The relationship was shown between the deformable method discussed earlier in this thesis and the rest of the segmentation framework. Both a template deformation and multiscale region analysis are combined to achieve a final classification.

Experiments were conducted to test the effectiveness of the template deformation within the framework. Comparison against manually segmented data shows that the deformed result is already quite good. The combined result did improve the result in some cases, but led to an overextension of the white matter in other cases. While the theory and experimental results provide hope that accurate segmentation can be automatically achieved, further research is still needed in this area.

CHAPTER 8

CONCLUSION

This thesis presented a systematic study of deformable image transformations. A method was proposed for nonrigidly aligning a template with a study image. This method is robust to a wide variety of contrast variations and supports large, curved deviations from the original template. The deformations are chosen in accordance with maximization of the mutual information between the image and template, a viscous fluid model, and learned shape information. Potential applications of the method include image segmentation, functional brain mapping, and automatic target recognition.

The thesis research contributes to the field of deformable image transformations in several ways. First, an in-depth study of mutual information was conducted, as it applies to nonrigid template deformation. A study of the bias and variance of the MI between images was done, and conditions for optimality of MI were defined. Limitations based on these conditions were investigated and ways of mitigating the impact of these limitations were presented. Templates with fewer graylevels and a distinct graylevel for each structure were more suitable for MI-based deformation. A formula was derived to convert MI to a body force. The calculation was found to scale more favorably with image size than previously reported calculations.

A bound on the maximal mutual information was derived. The bound can be calculated by maximizing the entropy of the template under the constraint that the conditional entropy of the template be zero, given the study image. Not only may the bound prove

useful in template initialization and selection, but it also provides insight into the usefulness of MI in a nonrigid setting. It was demonstrated that templates which achieve the bound may not give meaningful segmentations of the study image. Thus, the need for constraints on the deformation were clearly shown.

A fluid model was developed. The model maintains desirable properties of an existing fluid model while being computationally more efficient. The fluid partial differential equation models the Markov dependencies between template pixels. Experimental results showed that the fluid model allows large, curved deformations, but also allows over-deformation.

The shape model presented in Chapter 6 was shown to reduce the over-deformations inconsistent with prior knowledge about the shape of the skull and scalp, while still allowing large, nonlinear deformations in other parts of the template. The model captures both inter- and intrashape variations among distant parts of the template.

The proposed deformable template model was incorporated into a brain image segmentation method under development at the Beckman Institute for Advanced Science and Technology on the University of Illinois at Urbana-Champaign. The segmentation method was used to segment white and gray matter from brain images. Results were compared to manually segmented images. The results were mixed. Although the complete segmentation method improved the deformable template in some areas, other areas were degraded.

8.1 Areas of Further Research

Through the course of this research several possible avenues to pursue have been discovered; they are discussed here.

8.1.1 Incorporating other types of learning

Machine learning has been applied to the deformable procedure in the form a shape model trained by previously deformed templates. Other types of learning may also be

useful to the process. For example, as the deformation progresses through time, certain deformation sequences tend to occur. An associative memory may be able to retrieve the entire sequence from the first few steps.

8.1.2 Better integration of deformation and region information

Within the developmental segmentation framework, template deformation and multiscale region segmentation are done autonomously, then combined for the final result. Incremental inclusion of the region information into the deformation process may both improve the result and speed up the process. For example, deformations may be preferred which do not cross region boundaries. Resistance could be applied proportional to the number of scales for which boundaries are violated.

8.1.3 Hierarchical deformation

Research has been previously reported for different ways to formulate deformation procedures in a hierarchical fashion. Proposed methods include multiple resolutions of the image [24] and multiple degrees of fluidity or elasticity [46, 89]. Discussions in Sections 4.5 and 7.3 suggest another hierarchical formulation of the problem. A trade-off was found between the number of graylevels in the template and the stability of the deformation process. A possible modification of the process would deform to the large distinguishable structures first, such as white matter, gray matter, bone and skin; followed by smaller structures such as CSF or muscle.

8.1.4 Other data sets

Although there is no known limitation on the theory that would preclude application of the deformable model to higher than two dimensions, investigation may uncover difficulties associated with 3-D or higher deformations. Additionally, it would be interesting to apply the deformable method to other than medical images. The procedure could find

application in military or other commercial applications where image understanding is reliant on both prior information and image features.

APPENDIX A

IDENTITIES IN INFORMATION THEORY

The following are some properties of mutual information and entropy:

$$H(X) \geq 0 \quad (\text{A.1})$$

$$MI(X;Y) \geq 0 \quad (\text{A.2})$$

$$H(X | Y) \leq H(X) \quad (\text{A.3})$$

$$MI(X;Y) = MI(Y;X) \quad (\text{A.4})$$

$$MI(X;X) = H(X) \quad (\text{A.5})$$

$$MI(X;Y) \leq \min(H(X), H(Y)) \quad (\text{A.6})$$

$$MI(X;Y) \leq (H(X) + H(Y))/2 \quad (\text{A.7})$$

$$MI(X;Y) \leq \max(H(X), H(Y)) \quad (\text{A.8})$$

$$MI(X;Y) \leq H(X, Y) \quad (\text{A.9})$$

$$MI(X;Y) \leq H(X) + H(Y) \quad (\text{A.10})$$

$$MI(X;Y) = H(X) - H(X | Y) \quad (\text{A.11})$$

$$MI(X;Y) = H(Y) - H(Y | X) \quad (\text{A.12})$$

If X and Y are independent, some of the above inequalities turn into equalities:

$$MI(X;Y) = 0 \quad (\text{A.13})$$

$$H(X | Y) = H(X) \quad (\text{A.14})$$

Additionally, for independent X and Y , and a one-to-one function $F(\cdot)$

$$H(f(X) + Y \mid X) = H(X) \tag{A.15}$$

$$H(X + Y) \geq H(X) \tag{A.16}$$

$$H(X + Y) \geq H(Y) \tag{A.17}$$

APPENDIX B

DERIVATION OF THE FLUID MODEL COST FUNCTIONAL

The calculus of variations [90] states that $F(\mathbf{v})$ is stationary at $\mathbf{v}$ if

$$\frac{\partial}{\partial \epsilon} F(\mathbf{v} + \epsilon \eta) \Big|_{\epsilon \rightarrow 0} = 0 \quad (\text{B.1})$$

Let

$$\delta F \equiv \lim_{\epsilon \rightarrow 0} \frac{\partial}{\partial \epsilon} F(\mathbf{v} + \epsilon \eta)$$

Applying (B.1) to the first term on the right-hand side of (5.8) gives

$$\begin{aligned} F(\mathbf{v} + \epsilon \eta) &= \int |\nabla \mathbf{v} + \nabla \epsilon \eta|^2 d\mathbf{x} \\ &= \int (\nabla \mathbf{v} + \nabla \epsilon \eta)^T (\nabla \mathbf{v} + \nabla \epsilon \eta) d\mathbf{x} \\ &= \int (\nabla \mathbf{v})^T (\nabla \mathbf{v}) d\mathbf{x} + 2\epsilon \int (\nabla \mathbf{v})^T (\nabla \eta) + \epsilon^2 |\nabla \eta|^2 d\mathbf{x} \end{aligned} \quad (\text{B.2})$$

So the variation of F based on (B.1) is

$$\begin{aligned} \delta F &= 2 \int (\nabla \mathbf{v})^T (\nabla \eta) d\mathbf{x} \\ &= 2 \int \nabla \cdot \eta \nabla \mathbf{v} - \eta \nabla^2 \mathbf{v} d\mathbf{x} \end{aligned} \quad (\text{B.3})$$

The second line follows from the identity [91]

$$\nabla \cdot a \nabla b = \nabla a \cdot \nabla b + a \nabla^2 b$$

Using the divergence theorem leads to

$$\begin{aligned}\delta F &= 2 \oint_{\partial \mathcal{X}} \eta \nabla \mathbf{v} \cdot \hat{\mathbf{n}} d\mathbf{x} - 2 \int \eta \cdot \nabla^2 \mathbf{v} d\mathbf{x} \\ &= -2 \int \eta \cdot \nabla^2 \mathbf{v} d\mathbf{x}\end{aligned}\quad (\text{B.4})$$

where the surface integral is zero due to boundary conditions. Since we must ensure that B.4 goes to zero for arbitrary η , it must be that $\nabla \mathbf{v}(\mathbf{x}) = 0$

Similar derivation can be done for the second term of (5.8):

$$\begin{aligned}F(\mathbf{v} + \epsilon \eta) &= \int |\nabla \cdot \mathbf{v} + \nabla \cdot \epsilon \eta|^2 d\mathbf{x} \\ &= \int (\nabla \cdot \mathbf{v} + \nabla \cdot \epsilon \eta)^T (\nabla \cdot \mathbf{v} + \nabla \cdot \epsilon \eta) d\mathbf{x} \\ &= \int (\nabla \cdot \mathbf{v})^T (\nabla \cdot \mathbf{v}) d\mathbf{x} + 2\epsilon \int (\nabla \cdot \mathbf{v})^T (\nabla \cdot \eta) + \epsilon^2 |\nabla \cdot \eta|^2 d\mathbf{x}\end{aligned}\quad (\text{B.5})$$

So the variation of F based on (B.1) is

$$\delta F = 2 \int (\nabla \cdot \mathbf{v})^T (\nabla \cdot \eta) d\mathbf{x}\quad (\text{B.6})$$

which using identity

$$(\nabla \cdot \mathbf{v})(\nabla \cdot \eta) = \nabla \cdot (\eta \nabla \cdot \mathbf{v}) - \eta \cdot \nabla \nabla \cdot \mathbf{v}$$

becomes

$$\delta F = 2 \int \nabla \cdot (\eta \nabla \cdot \mathbf{v}) d\mathbf{x} - \int \eta \cdot \nabla \nabla \cdot \mathbf{v} d\mathbf{x}\quad (\text{B.7})$$

The second line follows from the identity [91]

$$\nabla \cdot a \nabla b = \nabla a \cdot \nabla b + a \nabla^2 b$$

Using the divergence theorem leads to

$$\delta F = 2 \oint_{\partial \mathcal{X}} (\nabla \cdot \mathbf{v}) \eta \cdot \hat{\mathbf{n}} d\mathbf{x} - 2 \int \eta \cdot \nabla \nabla \cdot \mathbf{v} d\mathbf{x}\quad (\text{B.8})$$

Again the surface integral goes to zero due to boundary conditions, and it must be that $\nabla \nabla \cdot \mathbf{v} = 0$ for the functional to be stationary.

REFERENCES

- [1] A. K. Jain, Y. Zhong, and M.-P. Dubuisson-Jolly, "Deformable template models: A review," *Sig. Proc.*, vol. 71, no. 2, pp. 109–129, 1998.
- [2] T. McInerney and D. Terzopoulos, "Deformable models in medical image analysis: A survey," *Medical Image Analysis*, vol. 1, no. 2, pp. 91–108, 1996.
- [3] B. Widrow, "The rubber mask technique," *Patt. Recog.*, vol. 5, no. 3, pp. 175–211, 1973.
- [4] M. Fischler and R. Elschlager, "The representation and matching of pictorial structures," *IEEE Trans. Computers*, vol. 22, no. 1, pp. 67–92, 1973.
- [5] J. Agris, R. deFigueiredo, G. Hillman, and T. Kent, "A novel method for 3-D segmentation and volume estimation of brain compartments from MRI," in *Proc. Ann. Int. Conf. IEEE Engineering in Medicine and Biology Society*, vol. 13, 1991, pp. 68–70.
- [6] X. Ji, S. Wang, M. B. Skouson, and Z.-P. Liang, "Brain image segmentation by multiscale analysis and template matching," in *Int. Soc. Magn. Reson. Med.*, 1999, p. 2186.
- [7] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, 1987.
- [8] C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow," *IEEE Trans. Image Proc.*, vol. 7, no. 3, pp. 359–369, 1998.
- [9] C. A. Davatzikos and J. L. Prince, "An active contour model for mapping the cortex," *IEEE Trans. Med. Imaging*, vol. 14, pp. 65–80, Mar. 1995.
- [10] S. Lobregt and M. A. Viergever, "A discrete dynamic contour model," *IEEE Trans. Med. Imaging*, vol. 14, pp. 12–24, Mar. 1995.
- [11] L. D. Cohen, "On active contour models and balloons," *CVGIP: Image Understanding*, vol. 53, pp. 211–218, Mar. 1991.
- [12] D. Geiger, A. Bupta, L. A. Costa, and J. Vlontzos, "Dynamic programming for detecting, tracking, and matching deformable contours," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 17, pp. 294–302, Mar. 1995.
- [13] T. McInerney and D. Terzopoulos, "Topologically adaptable snakes," in *Proc. Fifth Int. Conf. on Computer Vision (ICCV'95)*, June 1995, pp. 840–845.

- [14] A. J. Abrantes and J. S. Marques, "A class of constrained clustering algorithms for object boundary extraction," *IEEE Trans. Image Proc.*, vol. 5, no. 11, pp. 1507–1521, 1996.
- [15] C. A. Davatzikos, "Spatial transformation and registration of brain images using elastically deformable models," *Comp. Vision and Image Understanding*, vol. 66, pp. 207–222, May 1997.
- [16] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *Int. J. Comput. Vis.*, vol. 22, no. 1, pp. 61–79, 1997.
- [17] W. J. Niessen et al., "Geodesic deformable models for medical image analysis," *IEEE Trans. Med. Imaging*, vol. 17, pp. 634–641, Aug. 1998.
- [18] R. Kimmel, N. Kiryati, and A. M. Bruckstein, "Analysing and synthesizing images by evolving curves with the Osher-Sethian method," *Int. J. Comput. Vis.*, vol. 24, no. 1, pp. 37–55, 1997.
- [19] B. B. Kimia, A. R. Tannenbaum, and S. W. Zucker, "Shapes, shocks, and deformations I: The components of two-dimensional shape and the reaction diffusion space," *Int. J. Comput. Vis.*, vol. 15, pp. 189–224, 1995.
- [20] R. Malladi, J. A. Sethian, and B. C. Vemuri, "Shape modeling with front propagation: A level set approach," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 17, no. 2, pp. 158–175, 1995.
- [21] A. Yezzi, S. Kichenassamy, A. Kumar, P. Oliver, and A. Tannenbaum, "A geometric snake model for segmentation of medical imagery," *IEEE Trans. Med. Imaging*, vol. 16, pp. 199–209, 1997.
- [22] S. Osher and J. A. Sethian, "Fronts propagating with curvature dependent speed: Algorithms based on Hamiltonian-Jacobi formulation," *J. of Computational Physics*, vol. 79, pp. 12–49, 1988.
- [23] C. Broit, "Optimal registration of deformed images," Ph.D. dissertation, University of Pennsylvania, Philadelphia, PA, 1981.
- [24] R. Bajcsy and S. Kovacic, "Multiresolution elastic matching," *Comp. Vis. Graph. Image Proc.*, vol. 46, pp. 1–21, 1989.
- [25] G. E. Christensen, R. D. Rabbitt, and M. I. Miller, "Deformable templates using large deformation kinematics," *IEEE Trans. Image Proc.*, vol. 5, no. 10, pp. 1435–1447, 1996.
- [26] T. McInerney and R. Kikinis, "An object-based volumetric deformable atlas for the improved localization of neuroanatomy in MR images," in *Medical Image computing and computer-assisted Intervention – MICCAI'98*, 1998, pp. 861–869.
- [27] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," *Nature*, vol. 317, no. 6035, pp. 314–319, 1985.

- [28] N. J. I. Mars et al., "Time delay estimation in non-linear systems using average amount of mutual information analysis," *Signal Processing*, vol. 4, pp. 139–153, 1982.
- [29] P. Viola and W. W. Wells, "Alignment by maximization of mutual information," in *Proc. 5th Int. Conf. Computer Vision*, 1995, pp. 15–23.
- [30] W. M. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information," *Medical Image Analysis*, vol. 1, no. 1, pp. 25–51, 1996.
- [31] A. Collignon et al., "Automated multi-modality image registration based on information theory," in *Proc. Information Processing in Medical Imaging Conf.*, 1995, pp. 263–274.
- [32] F. Maes, A. Collignon, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imaging*, vol. 16, pp. 187–198, 1997.
- [33] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "Incorporating connected region labelling into automated image registration using mutual information," in *Proc. MMBIA*, 1996, pp. 23–31.
- [34] N. Hata, T. Dohi, S. Warfield, W. M. Wells, R. Kikinis, and F. A. Jolesz, "Multi-modality deformable registration of pre- and intraoperative images for MRI-guided brain surgery," in *Medical Image computing and computer-assisted Intervention – MICCAI'98*, 1998, pp. 1067–1074.
- [35] J. B. A. Maintz, E. H. W. Meijering, and M. A. Viergever, "General multimodal elastic registration based on mutual information," in *SPIE Medical Imaging 1998: Image Processing*, vol. 3338, 1998, pp. 144–154.
- [36] T. Gaens, F. Maes, D. Vandermuelen, and P. Suetens, "Non-rigid multimodal image registration using mutual information," in *Medical Image computing and computer-assisted Intervention – MICCAI'98*, 1998, pp. 1099–1106.
- [37] Y. G. Leclerc, "Constructing simple stable descriptions for image partitioning," *Int. J. Comput. Vis.*, vol. 3, pp. 73–102, 1989.
- [38] T. Kanungo, B. Dom, W. Niblack, and D. Steele, "A fast algorithm for MDL-based multi-band image segmentation," in *Comp. Vision Patt. Recog.*, 1994, pp. 609–616.
- [39] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam, "Use of active shape models for locating structures in medical images," *Image and Vision Comp.*, vol. 12, pp. 355–365, July/August 1994.
- [40] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models—their training and application," *Comp. Vision and Image Understanding*, vol. 61, pp. 38–59, Jan. 1995.

- [41] B. Szekely, A. Kelemen, C. Brechbuhler, and G. Gerig, "Segmentation of 3D objects from MRI volume data using constrained elastic deformations of flexible Fourier surface models," in *CVRMed'95*, 1995, pp. 495–505.
- [42] N. Duta and M. Sonka, "Segmentation and interpretation of MR brain images: An improved active shape model," *IEEE Trans. Med. Imaging*, vol. 17, pp. 1049–1062, Dec. 1998.
- [43] D. Terzopoulos and D. Metaxas, "Dynamic 3-D models with local and global deformations: Deformable superquadrics," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 13, pp. 703–714, July 1991.
- [44] B. C. Vemuri and A. Radisavljevic, "Multiresolution stochastic hybrid shape models with fractal priors," *ACM Transactions on Graphics*, vol. 13, pp. 177–120, Apr. 1994.
- [45] E. Bardinet, L. D. Cohen, and N. Ayache, "A parametric deformable model to fit unstructured 3D data," *Comp. Vision and Image Understanding*, vol. 71, pp. 39–54, July 1998.
- [46] G. E. Christensen, H. J. Johnson, J. W. Haller, J. Melloy, M. W. Vannier, and J. L. Marsh, "Synthesizing average 3D anatomical shapes using deformable templates," in *SPIE Medical Imaging 1999: Image Processing*, vol. 3661, 1999, pp. 574–582.
- [47] S. C. Joshi, "Large deformation diffeomorphisms and gaussian random fields for statistical characterization of brain sub-manifolds," Ph.D. dissertation, Washington University, Saint Louis, MO, 1998.
- [48] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [49] R. M. Gray and L. D. Davisson, *Random Processes, A Mathematical Approach for Engineers*. Englewood Cliffs, NJ: Prentice-Hall, 1986.
- [50] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.
- [51] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Englewood Cliffs, NJ: Prentice Hall, 1999.
- [52] H. Stark and J. W. Woods, *Probability, Random Processes, and Estimation Theory for Engineers*. Englewood Cliffs, NJ: Prentice Hall, 1994.
- [53] P. A. Viola, "Alignment by maximization of mutual information," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1995.
- [54] M. Rosenblatt, "Curve estimates," *Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1815–1842, 1971.
- [55] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*. New York: Wiley, 1980.
- [56] V. A. Epanechnikov, "Nonparametric estimates of a multivariate probability density," *Theor. Probability Appl.*, vol. 14, pp. 153–158, 1969.

- [57] M. B. Skouson and Z.-P. Liang, "Template deformation by maximizing mutual information," in *Proc. BMES/EMBS*, 1999, p. 1162.
- [58] F. Maes, "Segmentation and registration of multimodal medical images," Ph.D. dissertation, Katholieke Universiteit Leuven, Leuven, Belgium, 1998.
- [59] A. P. Zijdenbos and B. B. Dawant, "Brain segmentation and white matter lesion detection in MR images," *Critical Reviews in Biomedical Engineering*, vol. 22, no. 5/6, pp. 401-465, 1994.
- [60] R. Moddemeijer, "On estimation of entropy and mutual information on continuous distributions," *Signal Processing*, vol. 16, pp. 233-248, Mar. 1989.
- [61] G. E. Christensen, "Deformable shape models for anatomy," Ph.D. dissertation, Washington University, Saint Louis, MO, 1994.
- [62] R. C. Buck, *Advanced Calculus*. St. Louis, MO: McGraw-Hill, 1978.
- [63] E. W. Weisstein, *CRC Concise Encyclopedia of Mathematics*. Boca Raton, FL: CRC Press, 1999.
- [64] M. B. Skouson and M. E. Kowalski, "A simple and efficient method for image segmentation with deformable templates," in *IEEE Southwest Symposium on Image Analysis and Interpretation*, Apr. 2000, pp. 31-35.
- [65] M. T. Heath, *Scientific Computing: An Introductory Survey*. New York: McGraw-Hill, 1997.
- [66] M. Bro-Nielsen and C. Gramkow, "Fast fluid registration of medical images," in *Visualization in Biomedical Computing, 4th International Conference, VBC '96*, 1996, pp. 267-276.
- [67] M. Frigo and S. G. Johnson, "FFTW: An adaptive software architecture for the FFT," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98*, vol. 3, IEEE, 1998, pp. 1381-1384.
- [68] R. A. Johnson and D. W. Wichern, *Multivariate Statistics, A Practical Approach*. London/New York: Chapman and Hall, 1988.
- [69] N. Duta, A. K. Jain, and M.-P. Dubisson-Jolly, "Learning 2D shape models," in *Comp. Vision Patt. Recog.*, vol. 2, 1999, pp. 8-14.
- [70] D. Marr and E. C. Hildreth, "Theory of edge detection," in *Proc. Roy. Soc. London, B*, vol. 207, 1980, pp. 181-217.
- [71] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE Trans. Inf. Theory*, vol. 38, pp. 617-643, 1992.
- [72] S. W. Zucker, "Survey region growing: Childhood and adolescence," *Comp. Graph. and Image Proc.*, vol. 5, pp. 382-399, 1976.
- [73] S. Harowitz and T. Pavlidis, "Picture segmentation by a directed split-and-merge procedure," in *Proc. 2nd Int. Joint Conf. Pattern Recognition*, 1974, pp. 424-433.

- [74] R. M. Haralick and L. G. Shapiro, "Survey: Image segmentation techniques," *Comp. Vis. Graph. Image Proc.*, vol. 29, pp. 100-32, 1985.
- [75] K. Haris, S. N. Efstratiadis, and N. Maglaveras, "Watershed-based image segmentation with fast region merging," in *Int. Conf. Image Proc.*, vol. 3, 1998, pp. 338-342.
- [76] N. Ahuja, "A transform for multiscale image segmentation by integrated edge and region detection," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 18, pp. 1211-1235, 1996.
- [77] M. Tabb and N. Ahuja, "Multiscale image segmentation by integrated edge and region detection," *IEEE Trans. Image Proc.*, vol. 6, no. 5, pp. 642-654, 1997.
- [78] M. Sonka et al., "Knowledge-based interpretation of MR brain images," *IEEE Trans. Med. Imaging*, vol. 15, no. 4, pp. 443-452, 1998.
- [79] M. S. Brown et al., "Method for segmenting chest CT image data using an anatomical model: Preliminary results," *IEEE Trans. Med. Imaging*, vol. 16, no. 6, pp. 828-839, 1997.
- [80] S. P. Raya, "Low-level segmentation of 3D magnetic resonance brain images—a rule-based system," *IEEE Trans. Med. Imaging*, vol. 9, pp. 327-337, 1990.
- [81] K. Held et al., "Markov random field segmentation of brain MR images," *IEEE Trans. Med. Imaging*, vol. 16, no. 6, pp. 878-886, 1997.
- [82] C. A. Bouman and M. Shapiro, "A multiscale random field model for Bayesian image segmentation," *IEEE Trans. Image Proc.*, vol. 3, no. 2, pp. 166-177, 1994.
- [83] K. L. Vincken, A. S. E. Koster, and M. A. Viergever, "Probabilistic multiscale image segmentation," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 19, pp. 109-120, 1997.
- [84] Y. Salu and J. Tilton, "Classification of multispectral image data by binary diamond neural network and by nonparametric, pixel-by-pixel methods," *IEEE Tran. on Geoscience and Remote Sensing*, vol. 31, no. 3, pp. 606-616, 1993.
- [85] W. J. Niessen, K. L. Vincken, J. Weickert, and M. A. Viergever, "Three-dimensional MR brain segmentation," in *Sixth International Conf. on Computer Vision. Bombay, India*, Jan. 1998, pp. 53-58.
- [86] G. E. Christensen, S. C. Joshi, and M. L. Miller, "Volumetric transform of brain anatomy," *IEEE Trans. Med. Imaging*, vol. 16, pp. 864-877, Jan. 1997.
- [87] J. West et al., "Comparison and evaluation of retrospective intermodality brain image registration techniques," *J. Comput. Assist. Tomog.*, vol. 21, no. 4, pp. 554-566, 1997.
- [88] I. G. Zubal, C. R. Harrell, E. O. Smith, Z. Rattner, G. R. Gindi, and P. B. Hoffer, "Computerized three-dimensional segmented human anatomy," *Medical Physics*, vol. 21, pp. 199-302, February 1994.

- [89] G. E. Christensen, "Consistent linear-elastic transformations for image matching," in *LNCS*, vol. 1613, 1999, pp. 224–237.
- [90] A. P. Sage and I. W. Chelsea C., *Optimum Systems Control*. Englewood Cliffs, NJ: Prentic Hall, 1977.
- [91] W. C. Chew, *Waves and Fields in Inhomogeneous Media*. New York: IEEE Press, 1995.

VITA

Mark Brown Skouson was born in Massachusetts in November 1968. He received a B.S. degree in electrical and computer engineering from Brigham Young University in Provo, Utah, in 1993. In 1994, he also received his M.S. degree in electrical engineering from the same institution. His master's thesis was entitled, "Learning-Based Pattern Recognition Using Vector Quantization." During his studies at Brigham Young University, he was employed as a research assistant, teaching assistant, and recitation lecturer.

Mark Skouson accepted a commission in the United States Air Force in 1993 and was stationed at Holloman Air Force Base in New Mexico from 1994 to 1997, where he performed duties as a flight test engineer and radar cross section project engineer. His numerous honors included the Achievement Medal, Aerial Achievement Medal, and the Meritorious Service Medal. He left New Mexico in 1997 to pursue his Ph.D. at the University of Illinois. He is a member of the Institute of Electrical and Electronics Engineers (IEEE).

Selected Publications

M. B. Skouson and Z.-P. Liang, "Template deformation constrained by shape priors," in *Proc. CVPR*, June 2000.

M. B. Skouson and M. E. Kowalski, "A simple and efficient method for image segmentation with deformable templates," in *IEEE Southwest Symposium on Image Analysis and Interpretation* (Austin, Texas), Apr. 2000, pp. 31-35.

M. B. Skouson and Z.-P. Liang, "Template deformation by maximizing mutual information," in *Proc. BMES/EMBS* (Atlanta, GA), 1999, p. 1162.

X. Ji, S. Wang, M. B. Skouson, and Z.-P. Liang, "Brain image segmentation by multiscale analysis and template matching," in *Int. Soc. Magn. Reson. Med.* (Philadelphia, PA), 1999.