# BIBLIOGRAPHY

[1] A. Antonacopoulos. Page segmentation using the description of the background. *Computer Vision and Image Understanding*, pages 350–369, June 1998.

[2] H. S. Baird. Background structure in document images. *Document Image Analysis*, pages 17–34, 1994.

[3] D. Baker and A. K. McCallum. Distributional clustering of words for text classification. In *Proc. SIGIR'98*, pages 96–103, Melbourne, Australia, 1998.

[4] F. Bapst and R. Ingold. Using typography in document image analysis. *Electronic Publishing, Artistic Imaging, and Digital Typography. EP'98 & RIDT'98 Proceedings.*, pages 240–251, Mar./Apr. 1998.

[5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proc. ACM SIGGRAPH '99 Conference Proceedings*, pages 187–194, 1999.

[6] F. Le Bourgeois, H. Emptoz, and S. Souafi Bensafi. Document understanding using probabilistic relaxation: Application on tables of contents of periodicals. In *Sixth International Conference on Document Analysis and Recognition(ICDAR01)*, pages 508–512, Seattle, WA, September 2001.

[7] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *In International Conference on Computer Vision*, 1999.

[8] W. Buntine. Learning classification trees. *Statistics and Computing journal*, pages 63–76, 1992.

[9] M. Campione, K. Walrath, and A. Huml. *The Java(TM) Tutorial: A Short Course on the Basics(The Java(TM) Series).*

[10] F. Cesarini, M. Gori, S. Marinai, and G. Soda. Informys: A flexible invoice-like form-reader system. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(7):690–706, July 1998.

[11] F. Cesarini, M. Lastri, S. Marinai, and Giovanni Soda. Encoding of modified x-y trees for document classification. In *Sixth International Conference on Document Analysis and Recognition(ICDAR01)*, pages 1131–1135, Seattle, WA, September 2001.

[12] S. Chandran and R. Kasturi. Structural recognition of tabulated data. In *Proceedings of International Conference on Advances in Pattern Recognition (ICAPR) 93*, pages 516–519, Tsukuba Science City, Japan, October 1993.

[13] B. Chandrasekan and T. Harley. Comments on the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 15:421–423, 1969.

[14] C.-h. Chen. *Statistical Pattern Recognition.* Spartan Books, New Jersey, 1973.

[15] H.-H. Chen, S.-C. Tsai, and J.-H. Tsai. Mining tables from large scale html texts. In *The 18th International Conference on Computational Linguistics*, Saabrucken, Germany, July 2000.

[16] S. Chen, R. M. Haralick, and I. T. Phillips. Simultaneous word segmentation from document images using recursive morphological closing transform. *Proceedings of the 3rd ICDAR*, pages 761–764, Aug. 1995.

[17] D. Chetverikov, J. Liang, J. Komuves, and R. Haralick. Zone classification using texture features. In *Proc. International Conference on Pattern Recognition*, pages 676–680, Vienna, 1996.

[18] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–296, August 1995.

[19] F. Esposito, D. Malerba, and G. Semeraro. Classification in noisy environments using a distance measure between structural symbolic descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:390–402, March 1992.

[20] F. Esposito, D. Malerba, G. Semeraro, E. Annese, and G. Scafuro. An experimental page layout recognition system for office document automatic classification: An integrated approach for inductive generalization. In *Proceedings of the 10 th International Conference on Pattern Recognition*, pages 557 – 562, Atlantic City, 1990.

[21] K. C. Fan and L. S. Wang. Classification of document blocks using density features and connectivity histogram. *Pattern Recognition Letters*, 16:955–962, September 1995.

[22] B. Gatos, S. L. Mantzaris, and A. Antonacopoulos. First international newspaper segmentation contest. In *Sixth International Conference on Document Analysis and Recognition(ICDAR01)*, pages 1190–1194, Seattle, WA, September 2001.

[23] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Addison-Wesley, Reading, MA, USA, 1992.

[24] M. Goossens, F. Mittelbach, and A. Samarin. *The LaTeX Companion*. Addison-Wesley Publishing Company, 1994.

[25] E. Green and M. Krishnamoorthy. Model-based analysis of printed tables. In *Proceedings of the 3rd ICDAR*, pages 214–217, Canada, August 1995.

[26] D. Greig, B. Porteous, and A. Seheul. Exact maximum a posterior estimation for binary images. *Journal of the Royal Statistical Society, Series B*, pages 51(2):271–279, 1989.

[27] J. Ha, I. T. Phillips, and R. M. Haralick. Recursive x-y cut using bounding boxes of connected components. In *Proceedings of the Second International Conference on Document Analysis and Recognition*, pages 952–955, Tsukuba, Japan, October 1993.

[28] J. C. Handley. Table analysis for multi-line cell identification. In *SPIE Document Recognition and Retrieval VIII*, San Jose, California, January 2001.

[29] R. Haralick. Document image understanding: Geometric and logical layout. In *Proc. of the Conference on Computer Vision and Pattern Recognition*, pages 385–390, Seattle, WA, June 1994.

[30] R. Haralick and L. Shapiro. *Computer and Robot Vision*, volume 1. Addison Wesley, 1997.

[31] G. Harit, S. Chaudhury, P. Gupta, N. Vohra, and S. D. Joshi. A model guided document image analysis scheme. In *Sixth International Conference on Document Analysis and Recognition(ICDAR01)*, pages 1137–1141, Seattle, WA, September 2001.

[32] J. Hu, R. Kashi, D. Lopresti, G. Nagy, and G. Wilfong. Why table ground-truthing is hard. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, pages 129–133, Seattle WA, USA, September 2001.

[33] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong. Medium-independent table detection. In *SPIE Document Recognition and Retrieval VII*, pages 291–302, San Jose, California, January 2000.

[34] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong. Table structure recognition and its evaluation. In *SPIE Document Recognition and Retrieval VIII*, San Jose, California, January 2001.

[35] M. Hurst. Layout and language: Challenges for table understanding on the web. In *Web Document Analysis, Proceedings of the 1st International Workshop on Web Document Analysis*, pages 27–30, Seattle WA, USA, September 2001.

[36] RAF Technology Inc. *DAFS:Document Attribute Format Specification.* 1995.

[37] RAF Technology Inc. *Illuminator User's Manual.* 1995.

[38] D. J. Ittner and H. S. Baird. Language-free layout analysis. *Proceedings of the 2rd ICDAR*, pages 336–340, Oct. 1993.

[39] A. K. Jain, A. M. Namboodiri, and J. Subrahmonia. Structure in on-line documents. In *Sixth International Conference on Document Analysis and Recognition(ICDAR01)*, pages 844–848, Seattle, WA, September 2001.

[40] A.K. Jain and B. Yu. Document representation and its application to page decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:294–307, March 1998.

[41] N. Jin and Y. Y. Tang. Text area localization under complex-background using wavelet decomposition. In *Sixth International Conference on Document Analysis and Recognition(ICDAR01)*, pages 1126–1130, Seattle, WA, September 2001.

[42] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *Proc. 14th International Conference on Machine Learning*, pages 143–151, Morgan Kaufmann, 1997.

[43] A. Kam and G. Kopec. Document image decoding by heuristic search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:945–950, September 1996.

[44] T. Kanungo. Dvi2tiff user manual. *UW English Document Image Database - (I) Manual*, 1993.

[45] T. Kieninger and A. Dengel. Applying the t-rec table recognition system to the business letter domain. In *Sixth International Conference on Document Analysis and Recognition(ICDAR01)*, pages 518–522, Seattle, WA, September 2001.

[46] T. G. Kieninger. Table structure recognition based on robust block segmentation. *Document Recognition V.*, pages 22–32, January 1998.

[47] T. G. Kieninger and A. Dengel. A paper-to-html table converting system. In *Proceedings of Document Analysis Systems (DAS) 98*, Nagano, Japan, November 1998.

[48] T. G. Kieninger and A. Dengel. Table recognition and labeling using intrinsic layout features. In *Proceedings of International Conference on Advances in Pattern Recognition (ICAPR) 98*, Plymouth, UK, November 1998.

[49] K. Kise, A. Sato, and M. Iwata. Segmentation of page images using the area voronoi diagram. *Computer Vision and Image Understanding*, pages 370–382, June 1998.

[50] B. Klein, S. Gokkus, T. Kieninger, and A. Dengel. Three approaches to "industrial" table spotting. In *Sixth International Conference on Document Analysis and Recognition(ICDAR01)*, pages 513–517, Seattle, WA, September 2001.

[51] G. E. Kopec and P. A. Chou. Document image decoding using markov source models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16:602–617, June 1994.

[52] G. E. Kopec and M. Lomelin. Supervised template estimation for document image decoding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:1313–1324, December 1997.

[53] D. X. Le, J. Kim, G. Pearson, and G. R. Thom. Automated labeling of zones from scanned documents. *Proceedings SDIUT99*, pages 219–226, 1999.

[54] K. Lee, Y. Choy, and S. Cho. Geometric structure analysis of document images: A knowledge-based approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1224–1240, November 2000.

[55] J. Liang. *Document Structure Analysis and Performance Evaluation.* Ph.D thesis, Univ. of Washington, Seattle, WA, 1999.

[56] J. Liang, R. Haralick, and I. T. Phillips. Document zone classification using sizes of connected components. *Document Recognition III, SPIE'96*, pages 150–157, 1996.

[57] J. Liang, I. T. Phillips, and R. M. Haralick. Performance evaluation of document layout analysis algorithms on the uw data set. In *Document Recognition IV, Proceedings of the SPIE'97*, pages 149–160, San Jose, CA, 1997.

[58] J. Liang, I. T. Phillips, and R. M. Haralick. A statistically based, highly accurate text-line segmentation method. In *Proceedings of the 5th International Conference on Document Analysis and Recognition (ICDAR99)*, pages 551–554, Bangalore, India, September 1999.

[59] J. Liang, I. T. Phillips, and R. M. Haralick. A unified approach for document structure analysis and its application to text-line extraction. In *Proceedings of the 1999 Symposium on Document Image Understanding Technology (SDIUT99)*, pages 32–41, Annapolis, Maryland, April 1999.

[60] J. Liang, I. T. Phillips, and R. M. Haralick. Consistent partition and labeling of text blocks. *Journal of Pattern Analysis and Applications*, 3:196–208, 2000.

[61] J. Liang, I. T. Phillips, and R. M. Haralick. A methodology for document image structures extraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:719–734, July 2001.

[62] J. Liang, R. Rogers, R. M. Haralick, and I. T. Phillips. Uw-isl document image analysis toolbox: An experimental environment. In *Proceedings of the International Conference on Document Analysis and Recognition(ICDAR) '97*, pages 984–988, Ulm, Germany, August 1997.

[63] G. Liu and R. M. Haralick. Flir atr using location uncertainty. *Journal of Electronic Imaging*, 9:178–193, April 2000.

[64] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet portals with machine learning. In *Information Retrieval Journal*, volume 3, pages 127–163, Kluwer, 2000.

[65] T. Merz. *PostScript & Acrobat/PDF, Applications, Troubleshooting, and Cross-Platform Publishing.* Springer, 1997.

[66] T. P. Minka, D. S. Bloomberg, and K. Popat. Document image decoding using iterated complete path search. In *Document Recognition VIII*, San Jose CA, January 2001.

[67] P E. Mitchell and H. Yan. Newspaper document analysis featuring connected line segmentation. In *Sixth International Conference on Document Analysis and Recognition(ICDAR01)*, pages 1181–1185, Seattle, WA, September 2001.

[68] T. M. Mitchell. *Machine Learning.* McGraw-Hill, 1997.

[69] D. Mladenic. Text-learning and related intelligent agents. *IEEE Expert special issue on Applications of Intelligent Information Retrieval*, July-August 1999.

[70] A. Moffat and J. Zobel. Information retrieval systems for large document collections. In *The Third Text Retrieval Conference (TREC3)*, Gaithersburg, MD. National Institute of Standards and Technology, Special Publication 500225.

[71] G. Nagy. Twenty years of document image analysis in pami. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:38–62, January 2000.

[72] H. T. Ng, C. Y. Lim, and J. L. Koo. Learning to recognition tables in free text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 443–450, College Park, Maryland, 1999.

[73] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing.* Prentice Hall, Englewood Cliffs, NJ, USA, 1989.

[74] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on SMC, Vol. SMC-9*, pages 62–66, 1979.

[75] G. Penn, J. Hu, H. Luo, and R. McDonald. Flexible web document analysis for delivery to narrow-bandwidth devices. In *International Conference on Document Analysis and Recognition (ICDAR01)*, pages 1074–1078, Seattle WA, USA, September 2001.

[76] I. Phillips. Users' reference manual. *CD-ROM, UW-III Document Image Database-III*, 1995.

[77] I. Phillips, S. Chen, and R. Haralick. Cd-rom document database standard. In *Proceedings of the Second Intl. Conf. on Document Analysis and Recognition*, pages 478–483, Tsukuba Science City, Japan, October 1993.

[78] K. Popat, D. Bloomberg, and D. Greene. Adding linguistic constraints to document image decoding. *In Proceedings of the 4th IAPR Workshop on Document Analysis Systems (DAS2000)*, December 2000.

[79] K. Popat, D. Greene, J. K. Romberg, and D. Bloomberg. Adding linguistic constraints to document image decoding: Comparing the iterated complete path and stack algorithms. In *Document Recognition VIII*, San Jose CA, January 2001.

[80] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[81] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. *Numerical Recipes in C*. Cambridge University Press, 1988.

[82] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–285, February 1989.

[83] M. Rahgozar and R. Cooperman. A graph-based table recognition system. In *Document Recognition III, SPIE*, pages 192–203, San Jose, California, 1996.

[84] K. Sall. Xml: Structuring data for the web: An introduction. *http://wdvl.com/Authoring/Languages/XML/Intro*.

[85] J. H. Shamilian, H. S. Baird, and T. L. Wood. A retargetable table reader. In *Proceedings of the 4th ICDAR*, pages 158–163, Germany, August 1997.

[86] R. Sivaramakrishnan, I. Phillips, J. Ha, S. Subramanium, and R. Haralick. Zone classification in a document using the method of feature vector generation. In *Proceedings of the 3rd ICDAR*, pages 541–544, Montrael, Canada, August 1995.

[87] Inc. Texterity. *http://www.texterity.com*.

[88] T. A. Tokuyasu and P. A. Chou. An iterative approach to document image analysis. In *DLIA99 workshop*, Bangalore, India, September 1999.

[89] T. A. Tokuyasu and P. A. Chou. Turbo recognition:a statistical approach to layout analysis. In *Document Recognition VIII*, San Jose, CA, January 2001.

[90] V. N. Vapnik. *The Nature of Statistical Learning Theory*, volume 1. Springer, New York, 1995.

[91] Y. Wang. http://isl.wtc.washington.edu/ylwang/auttabgen.html.

[92] Y. Wang, R. Haralick, and I. T. Phillips. Improvement of zone content classification by using background analysis. In *Fourth IAPR International Workshop on Document Analysis Systems. (DAS2000)*, Rio de Janeiro, Brazil, December 2000.

[93] Y. Wang, R. Haralick, and I. T. Phillips. Zone content classification and its performance evaluation. In *Sixth International Conference on Document Analysis and Recognition(ICDAR01)*, pages 540–544, Seattle, WA, September 2001.

[94] Y. Wang and J. Hu. Detecting tables in html documents. In *Fifth IAPR International Workshop on Document Analysis Systems*, Princeton, New Jersey, USA, August 2002.

[95] Y. Wang and J. Hu. A machine learning based approach for table detection on the web. In *The Eleventh International World Web Conference*, Honolulu, Hawaii, USA, May 2002.

[96] Y. Wang, I. T. Phillips, and R. Haralick. From image to sgml/xml representation: One method. In *International Workshop on Document Layout Interpretation and Its Applications (DLIA'99)*, Bangalore, India, September 1999.

[97] Y. Wang, I. T. Phillips, and R. Haralick. Statistical-based approach to word segmentation. In *15th International Conference on Pattern Recognition, ICPR2000*, volume 4, pages 555–558, Barcelona, Spain, September 2000.

[98] Y. Wang, I. T. Phillips, and R. Haralick. Automatic table ground truth generation and a background-analysis-based table structure extraction method. In *Sixth International Conference on Document Analysis and Recognition(ICDAR01)*, pages 528–532, Seattle, WA, September 2001.

[99] Y. Wang, I. T. Phillips, and R. Haralick. Random table and its ground truth automatic generation: A tool for table understanding research. In *International Workshop on Document Layout Interpretation and Its Applications (DLIA'2001)*, Seattle, WA, USA, September 2001.

[100] Y. Wang, I. T. Phillips, and R. Haralick. Document zone content classification using decision tree and hmm. In *International Conference on Pattern Recognition*, Quebec City, Canada, August 2002.

[101] Y. Wang, I. T. Phillips, and R. Haralick. A study on the document zone content classification problem. In *Fifth IAPR International Workshop on Document Analysis Systems*, Princeton, New Jersey, USA, August 2002.

[102] Y. Wang, I. T. Phillips, and R. Haralick. Table detection via probability optimization. In *Fifth IAPR International Workshop on Document Analysis Systems*, Princeton, New Jersey, USA, August 2002.

[103] T. Watanabe, Q. Luo, and N. Sugie. Layout recognition of multi-kinds of table-form documents. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(4):432–445, April 1995.

[104] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Prof. SIGIR'99*, pages 42–49, Berkeley, California, USA, August 1999.

[105] M. Yoshida, K. Torisawa, and J. Tsujii. A method to integrate tables of the world wide web. In *Web Document Analysis, Proceedings of the 1st International Workshop on Web Document Analysis*, pages 31–34, Seattle WA, USA, September 2001.

[106] B. Yu and A. K. Jain. A generic system for form dropout. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(11):1,127–1,134, November 1996.

[107] K. Zuyev. Table image segmentation. In *Proceedings of the International Conference on Document Analysis and Recognition(ICDAR) '97*, pages 705–708, Ulm, Germany, August 1997.

# VITA

YALIN WANG

Department of Electrical Engineering, Box 352500

University of Washington

Seattle, WA 98195, U.S.A.

ylwang@u.washington.edu

## Academic Interests

Image processing, pattern recognition, document image analysis, and information retrieval.

## Education

- 1997-2002 University of Washington, Seattle, Washington, Ph.D in Electrical Engineering.
- 1994-1996 Tsinghua University, Beijing, China, M.S. in Computer Science.
- 1989-1994 Tsinghua University, Beijing, China, B.S. in Computer Science.

## Experience

- September 2001-March 2002 Queens College, City University of New York, Flushing, New York, Research Intern.
- June-September 2001 Avaya Labs Research, Basking Riedge, New Jersey, Summer Intern.

- 1997-2001 Intelligent Systems Laboratory, University of Washington, Seattle, Washington, Research assistant.

- December 1996-July 1997 Tsinghua University, Beijing, China, Associate Lecturer.

- July 1995-September 1996 Tsinghua Information Technology Company (QH-SOFT), Beijing, China, Senior System Analyst and Designer.

**Publications**

1. "Detecting Tables in HTML Documents", Y. Wang and J. Hu, Accepted by Fifty IAPR International Workshop on Document Analysis Systems (DAS2002), Princeton, NJ, USA, Aug. 2002.

2. "Table Detection Via Probability Optimization", Y. Wang, I.T. Phillips and R.M. Haralick, Accepted by Fifth IAPR International Workshop on Document Analysis Systems (DAS2002), Princeton, NJ, USA, Aug. 2002.

3. " A Study on the Document Zone Content Classification Problem", Y. Wang, I.T. Phillips and R.M. Haralick, Accepted by Fifth IAPR International Workshop on Document Analysis Systems (DAS2002), Princeton, NJ, USA, Aug. 2002

4. "Document Zone Content Classification Using Decision Tree and HMM", Y. Wang, I.T. Phillips and R.M. Haralick, Accepted by International Conference on Pattern Recognition, Quebec City, Canada, Aug. 11-15, 2002.

5. " A Machine Learning Based Approach for Table Detection on The Web", Y. Wang and J. Hu, Accepted by The Eleventh International World Web Conference, Honolulu, Hawaii, USA, May 7-11, 2002.

6. "Automatic Table Ground Truth Generation and A Background-analysis-based Table Structure Extraction Method", Y. Wang, I.T. Phillips and R.M. Haralick, Sixth International Conference on Document Analysis and Recognition , pp. 528-532, Seattle, Washington, U.S.A. Sep. 10-13, 2001.

7. "Zone Content Classification and Its Performance Evaluation", Y. Wang, R.M. Haralick and I.T. Phillips, Sixth International Conference on Document Analysis and Recognition , pp. 540-544, Seattle, Washington, U.S.A. Sep. 10-13, 2001

8. "Using Area Voronoi Tessellation to Segment Characters Connected to Graphics", Y. Wang, I.T. Phillips and R.M. Haralick, Workshop on Graphics Recognition (GREC2001), Kingston, Ontario, Canada. Sep. 2001

9. "Random Table and Its Ground Truth Automatic Generation: A Tool for Table Understanding Research", Y. Wang, I.T. Phillips and R.M. Haralick, Workshop on Document Layout Interpretation and its Applications (DLIA2001), Seattle, Washington, U.S.A. Sep. 2001

10. "Improvement of Zone Content Classification by Using Background Analysis", Y. Wang, R.M. Haralick and I.T. Phillips, Fourth IAPR International Workshop on Document Analysis Systems, Rio de Janeiro, Brazil, 10-13 December, 2000

11. "Statistical-based Approach to Word Segmentation", Y. Wang, I.T. Phillips and R.M. Haralick, 15th International Conference on Pattern Recognition, ICPR2000, Vol. 4, pp.555-558, Barcelona, Spain, September 2000.

12. "From Image to SGML/XML Representation: One Method", Y. Wang, I.T. Phillips and R.M. Haralick, International Workshop on Document Layout Interpretation and Its Applications (DLIA'99), Bangalore, India, September 1999

13. "A Heuristic Method to Recognize Dimension Sets in Architectural Engineering Drawings", Y. Wang, L. Tang and Z. Liu, Journal of Computer Science and Technology, Vol. 13, suppl. Issue; Dec. 1998, pp. 94-100

14. "A New Method to Recognize Dimension Set and its Application in Architectural Engineering Drawings", Y. Wang, L. Tang etc., Fifth International Conference on Computer Aided Design and Computer Graphics, Dec. 2-5, 1997, Shenzhen China