G.Nagy, Estimation, Learning and Adaptation: Systems that Improve with Use, Pierre DeVijver Award presentation, Procs. Structural, Syntactic and Statistical Pattern Recognition (SSSPR 2012), Hiroshima, Nov. 2012

ESTIMATION, LEARNING, AND ADAPTATION: SYSTEMS THAT IMPROVE WITH USE

George Nagy

RPI, Troy, NY 12180, USA

ABSTRACT. The accuracy of automated classification (labeling) of single patterns, especially printed, hand-printed, or handwritten characters, has leveled off. Further gains in accuracy require classifying sequences of patterns. Linguistic context, already widely used, relies on 1-D lexical and syntactic constraints. Style-constrained classification exploits the shape-similarity of sets of same-source (isogenous) characters of either the same or different classes. For understanding tables and forms, 2-D structural and relational constraints are necessary. Applications of pattern recognition that do not exceed the limits of human senses and cognition can benefit from green interaction wherein operator corrections are recycled to the classifier.

Keywords: Devijver, adaptive classification, style consistency, tables, green interaction

1 INTRODUCTION

Pierre Devijver and I shared several interests – nearest neighbors, Delaunay triangulation, clustering, connected components, error estimation, and context. I have dabbled in computational geometry, computer-aided design, remote sensing, and geographical information systems. However, most of my studies – and those of my students – have been devoted to document image analysis and to one of its fundamental components, character recognition.

The fact that research on character recognition has contributed so much to pattern recognition and machine learning cannot be attributed mainly to our desire to live in a paperless world. Character recognition is a limitless field of research in SPR because of the wealth of relationships induced by messages conveyed through sequences of visually recognizable patterns characterized by multi-dimensional feature vectors to be classified into possibly hierarchical classes with minimum error or cost. One can investigate and model the statistical distributions of individual features, of all the features of a single sample, and the relationships between the features of multiple patterns and class variables. Any of the above patterns may consist of a single stroke, a single letter or numeral, part of a word, or a whole word, page, or document. The various models used or proposed in character recognition can be represented concisely by Bayesian networks [1].

If this paper fails to conform to some rules, I claim the Senior Citizens' Exemption. First I shall reminisce about Pierre Devijver and his technical legacy. Then I shall ruminate about some of my own hobby horses, including a few whose connection with statistical pattern recognition may be less than obvious. Table processing, for example, may fit better under "syntactic and structural." The conclusions mention some developments that I did not anticipate.

The reader will find no formulas or experimental results herein. Since all of what I recount has already been published – some of it more than once – I keep to a bird's eye view. Details can be found in the references cited, or in the references cited in those references.

2 PIERRE DEVIJVER

Pierre Devijver was a man of many accomplishments, and I am honored to have a chance to talk about my favorite topic, statistical pattern recognition, under his aegis. I had occasion to meet Pierre at several conferences where our similar technical interests and a predilection for lunch-time walks promoted conversation. (If I had known that he was a marathon runner I would have worried about those long walks with him.)

I missed the 1973 ICPR in Washington DC where Pierre first introduced his ideas about the relationship of the Bayes Risk to the Mean Square Error [2]. He published an article on error bounds the following year in the *IEEE Transactions on Computers* (where the best papers on pattern recognition were published before *PAMI* came along) [3]. So even before I ever met him I had studied some of his work. Later, when I taught pattern recognition and document image processing at RPI, I benefited a great deal from his and Joseph Kittler's rigorous text [4], which includes some of their results on the bias and variance of k-NN based error estimates. I still return to their lucid presentation of probabilistic distance measures.

I know that Pierre and I interacted at the 1980 Pattern Recognition in Practice workshop organized by Edzard Gelsema† and Laveen Kanal in Amsterdam. We both presented papers at the ICPR earlier that year in Miami Beach, but I don't recall any conversation between us. I also cannot remember any specific discussion at the 1984 ICPR in Montreal. By then Pierre was vice-president of IAPR, so he must have spent any time left after the technical sessions at committee meetings. In spite of his conscientious work for IAPR, the following examples show that he found time for remarkable technical contributions.

2.1 Connected Components

For many years the Devijver-Ronse monograph on Connected Component Detection was the only book on the subject. It expounded efficient disk access while tracing a CC [5]. Pierre foresaw that CC detection would be a cornerstone of document image analysis even though at the time only a small swath of a scanned page would fit into primary memory.

2.2 Markov Algorithms

In 1966, working under first-order Markov Chain assumptions, Joseph Raviv devised an iterative algorithm to convert the information from feature vectors of preceding patterns into the prior probability for the current pattern by using bigram and trigram class-transition probability tables. At the 1984 ICPR, Pierre extended this to take into account the information from any number of *succeeding* patterns by adding an iterative backward pass [6,]. He subsequently improved the numerical stability of Baum's HMM training algorithm by computing joint rather than conditional probabilities [7].

Already then his interest lay in Markov meshes and MRFs for image processing, so he avoided any assumptions that applied only to 1-D. I wish that Pierre had completed his studies of MRFs a few years earlier, when my UNL students and I were struggling to set appropriate constraints on causal generation of 2-D Markov fields. Pierre's deep insights would have been invaluable.

2.3 Nearest Neighbors

As Godfried Toussaint repeatedly demonstrated, statistical pattern recognition and computational geometry mix well. The decision boundaries of a nearest neighbor classifier are a subset of the edges of its Voronoi diagram in feature space. The Voronoi diagram, in turn, can be computed rapidly from its dual Delaunay triangulation. One way to speed up nearest-neighbor classification is by removing from the reference set all the patterns with same-class Voronoi Neighbors. However, removing a pattern changes both the Voronoi diagram and the underlying Delaunay triangulation. Pierre devised fast algorithms for *dynamic* Delaunay triangulation in high-dimensions [8]. He also derived bounds on the fraction of the training set that can be harmlessly edited out.

Pierre's research included methods of estimating the error rate on the test set from results on the training set (still a hot topic), feature extraction, the relationship between clustering and mixture identification, and applications ranging from tumor detection to astronomy. This sketch of his contributions is far too superficial to do justice to his pervasive and persuasive ideas.

3 MY OWN TRAIL

Before I get into even a modicum of technicalities, I wish to acknowledge what good fortune I had in my collaborators and co-authors, and how much I learned from my students (some regrettably already retired). One of the best things about the field of pattern recognition is that it has attracted such bright, generous, and convivial scholars.

3.1 Feature Extraction

My first graduate student, at the Université de Montréal, was Kamal Abdali. I set him to solve the optimal feature extraction problem because nobody knew yet about NP-completeness. My last student, Xiaoli Zhang, confirmed my belief that the class-conditional statistical dependence structure (e.g., the covariance matrix) of features depends far more on the chosen feature set than on the data itself [9]. NP-complete it may be, but there is more to be done.

3.2 Unsupervised Classification

There is no such thing as unsupervised learning. Children learn without being explicitly taught, but only because they emulate the behavior of respected teachers (grown-up or other children). They often get to surpass the abilities of those whom they imitate. In 1965 we programmed a classifier to trust the labels assigned to scanned printed characters by an off-the-shelf journeyman classifier, and to use them for its own training set. On the data that it was re-trained on, this apprentice classifier turned out to be better than the original classifier, and so we used it as a role model for still another classifier. To our surprise and delight, the error rate kept dropping during several iterations and then flattened out [10,11]. Almost three decades later, these results impressed Henry Baird, so we tried "mean adaptation" with his own features and his then humungous 100-font dataset [12]. Henry concluded that the expected gain is considerable, while the downside risk was small.

Ho and Agrawala had pointed out earlier that we were lucky because the many datasets on which we had experimented all fell under restrictive constraints [13,14]. With features crafted by Hiromichi Fujisawa and Cheng-Lin Liu of the Hitachi Central Research Laboratory, Harsha Veeramachaneni used Expectation Maximization to re-estimate both the means and the class-covariance matrices using classifier-assigned labels. This turned out even luckier (at least on NIST hand-printed digits) than just mean adaptation [15].

We have not yet found necessary and sufficient conditions that would guarantee that adaptation will reduce the error on a set of same-source samples. Is

there a principled way to predict the results of adaptation? A place to start might be Castelli's and Cover's insights on the relative values of labeled and unlabeled samples.

3.3 Prototype Based Text-Image Compression

Any clustering method can be viewed as data compression with each cluster prototype serving as surrogate for all the patterns in that cluster. In 1970 or thereabouts, Pete Welch, my boss's boss at IBM Research, suggested that we apply the bitmap clustering methods that we had developed earlier for Chinese character recognition [16] to image data compression. It worked like a charm! We could not patent it because of a Government anti-monopoly suit, so IBM waited four years before letting us publish it [17].

Our method eventually resurfaced in DjVu and JBIG, but with a critical improvement that we had missed. We transmitted only run-length coded prototypes, or their (compressed) id and position. Subsequent researchers encoded the difference between the prototype and the actual glyph, thereby rendering the scheme lossless. Current methods are nearing the theoretical limits.

3.4 Decision Trees

Although I worked at IBM on several OCR projects, including the three million dollar reader for the Social Security Administration, the only algorithm that made it into a product was a probabilistic decision tree for isolated bitmapped characters [18]. Dick Casey developed most of the theory, and I programmed it up in APL during a summer at the IBM San Jose Research Center. After everything was reprogrammed efficiently in Japan it became the IBM TextReader. I still have copies of the shrink-wrapped floppies.

3.5 Language context

Both children and adults expand their vocabulary by guessing and refining the meaning of unknown words or phrases according to what makes sense time after time. If in a foreign land most street signs end in a particular string, it is likely to mean "street" or "avenue". Early proponents of the use of language context in pattern recognition include Allen Hanson, Ed Riseman†, Joe Raviv†, Godfried Toussaint, and Ching Suen.

Meanings can be assigned to unknown alphabetic glyphs so that they form words that are part of the language. Substitution ciphers have been solved this way since at least the days of the Roman Empire. I have participated in three initiatives to automate this process and apply it to scanned text.

In a first attempt, Dick Casey and I clustered bitmaps of scanned single-case English text in one of four different typefaces. We solved the resulting cryptograms by matching the frequencies of cluster numbers and bigrams of cluster numbers to the known letter unigram and bigram frequencies of the English language [19]. We were very pleased when Scientific American asked us to describe our methods in laymen's terms [20]. Twenty years later at the University of Nebraska, we improved the scheme by recursive matching of trial assignments against a lexicon of a few hundred words instead of letter n-grams [21]. In another ten years, Tin Ho at Bell Labs used a larger lexicon and improved the matching scheme. We demonstrated "OCR with no shape training" on Spitz glyphs at the Barcelona ICPR [22].

3.6 Style

At the 1992 ICPR I proposed exploiting the family resemblance of same-font letters and numerals for recognizing individually ambiguous characters when they appeared in their usual company [23]. I called this notion *spatial context.* During his doctoral research Prateek Sarkar dubbed the distinction between feature distributions originating from patterns from isogenous typeface, printer, writer or speaker, and distributions from patterns from heterogeneous sources, as *style*. Harsha Veeramachaneni explained style as follows: "the way Alice writes 1 helps predict the way she will write 7." Applying these concepts to fields of same-source patterns may not be so difficult, but defining them formally requires a lot of notation and subscripts.

A critical property of style context defined by Harsha is *order independence*, known in probability theory as *exchangeability* [24]. Pure style implies that the probability of any pattern field given the field class is equal to the probability of any permutation of that pattern field given the field class subjected to the same permutation. Order independence vitiates most types of language context, but style and multi-pattern language context can still be combined.

Another useful distinction can be made between *intra-class style* and *inter-class style.* In intra-class style, an "**e**" in a field of patterns to be recognized is always in the same font, or was written by the same person. One might, however, find this "**e**" next to a "**c**" of a different font or by a different writer. So there is some statistical dependence between the feature distributions of all the patterns of the same class, but no way to tell anything about a pattern of class *E* from a pattern of class *C*. The inter-class constraint is more rigid: samples from all the classes must be isogenous. Therefore the features of patterns even of different classes exhibit observable statistical dependence. Most of the adaptive classifiers discussed above require only intra-class style. The style classifiers described below make use of inter-class style.

Prateek Sarkar derived algorithms for optimal classification of style-consistent fields of arbitrary length [25]. He posited that the features of each pattern, while dependent on the features of other patterns in the field because of the same-style constraint, were independent of the *classes* of the other patterns. In other words, every "**e**" in the field looks the same, regardless of whether the field spells "element" or "dependent". He formulated several ways of combining Gaussians as mixture distributions to model the class-and-style-conditional probabilities via weighting factors that depend on both class and style. In terms of hand print, his method can classify fields of never-before-seen hybrid Ann-Jen script after training only on separate fields of Ann's, Bonnie's, Dave's and Jen's writing.

Because the computation of the optimal maximum likelihood assignment requires lengthy sum-of-products-of-sums computations, Prateek devised a *top-label* approximation equivalent to selecting from a set of style-specific feature classifiers the one that yields the highest field-feature likelihood. He trained his classifiers with a mixture of isogenous (*isofont*) fields, and tested them on isogenous fields of lengths different from those of the training set.

Harsha Veeramachaneni considered a continuous distribution of Gaussians instead of mixtures of a predetermined fixed number. His insight was that the posterior distribution of a field of any length can be determined from the cross-covariance matrices of only *pairs* of same-source pattern feature vectors. This led to quadratic field classification with computation proportional rather than exponential with field length [26].

From the perspective of style-constrained field classification, the field of an adaptive classifier encompasses the entire set of isogenous data rather than a fixed number of patterns. This observation may explain why some adaptive classifiers exploit only intra-class consistency. On short fields, on the other hand, more powerful and more computation-intensive classifiers can take full benefit of inter-class style consistency.

In practical OCR applications, style-constrained classification aims at scenarios similar to font or writer recognition. Both of these are effective tools for decreasing the error rate by substituting a single-font or single-writer classifier for a more error-prone omnifont or omni-writer classifier. In theory, however, style classifiers should achieve a lower error rate because they do not "waste" any statistical information on font or writer identification. In some applications, however, it is desirable to identify font or writer in addition to producing a transcription. We have also pursued, with mixed, success, variations of style classification, based on nearest-neighbors and support-vector machines.

4 TABLES

We began our studies of tables twenty years ago with foreign language tables that gave us a chance to see how much information can be derived from table structure without lexical help. Since then mainstream table recognition has progressed from scanned paper tables to computer-generated HTML and PDF tables. All of this work has been part of a long-standing and most enjoyable collaboration with Dave Embley (BYU), Sharad Seth (UNL), Moorthy Krishnamoorthy (RPI) and Dan Lopresti (Lehigh), often under the aegis of TANGO [27]. We have written far too many surveys and reports, especially considering how often our views have shifted, so rather than reciting progressive steps I just list some articles of faith (for which I take sole responsibility and which I may retract next year).

- The underlying grid of a table reveals a 2-D indexing scheme. This geometric indexing is interwoven with possibly higher-dimensional, logical "Wang" categories which can be interpreted as geometric indexing in a higher-dimensional space.

- The essential task of table analysis is to establish the relationship of column and row headers to individual data cells. This is trickier than might first appear because of the possible occurrence of hierarchical headers, spanning cells and headers in the row stub, and because the appearance of a table depends on the rendering program as well as the file containing the table Additional tasks require extracting metadata (table caption, title, footnote references, footnotes, aggregates, units, …).

- Tables are distinguished from *forms* because tables are meant to disseminate information rather than collect it. The distinction is often obvious, but a filled-out spreadsheet might be either a table or a form. In most forms individual field captions take the place of 2-D indexing. Their structure can be represented by graphs.

- Tables are distinguished from *lists* by 2-D indexing of data cells by row and column headers.. Even ordered lists like telephone books require a search to locate a cell. The table vs. list question arises only when nested lists of uniform length are laid out on a grid, or when table ill-formed table headers preclude unique indexing.

- Tables prepared for human readers are different from relational tables. The designer of a relational table must determine what is an attribute and what is a key, and orient the table accordingly, with attributes on top. In contrast, the orientation of tables prepared for hardcopy or web publication is usually determined by matching the number of row and column headers to the page or display format. Therefore in tables of Canadian statistics the column headers are often provinces, while in tables of US statistics the states are usually row headers. Visual tables are essentially symmetric with respect to rows and columns, but relational tables are not. This does not preclude the transformation of visual tables into relational tables.

5 GREEN INTERACTION

CAVIAR (*Computer Assisted Visual InterActive Recognition*) for flowers is an attempt at efficient human-computer interaction [28]. When a flower image is presented, the program extracts visually verifiable features (like the shape or number of petals), and classifies the flower into one of a hundred or so classes. If the user is unsatisfied, she or he can edit the features and reject or approve the classification according to the resemblance of the flower to reference samples. The classifier, in turn, adapts its parameters using what it learns from the user. Both the computer and the user improve with time. A most enjoyable part of this project was collecting over 600 wildflower samples.

In OCR, a perennial problem is obtaining large-enough *labeled* training sets. Studies have shown that classification on the test set improves even after tens or hundreds of thousands of training samples. The output of an OCR system, especially in error intolerant applications like medical or financial form entry, is often routed to operators for verification or correction. The best possible training set is the stream of data encountered during actual operation. Therefore all final corrected labels should be associated with the scanned patterns and routed back periodically to retrain the classifier.

Green interaction means that expensive and time-consuming human effort devoted to approving or correcting the output of any pattern recognition system should not be wasted. More on this at ICPR 2012.

6 CONCLUSION

We miss Pierre Devijver. We are fortunate that he left behind so much to think about.

Claims made ever since Pierre and I were starting out, to the effect that OCR was essentially a solved problem, turned out to be uninformed. I had the good luck to work on a variety of related problems with perspicacious colleagues and students. The study of each problem revealed new problems begging for solution. Writing surveys has shown me many more. It is like a Garden of Eden where the quandary is which fruit to taste first. Fortunately there are still some sweet low-hanging fruit left.

Progress in some areas surprises me. In 1968, when we worked on hand-printed numeral recognition, I was sure that writer-independent cursive script recognition was a pipe dream. Speech and face recognition also work better

than I expected. I underestimated the scope and power of Expectation Maximization (but eventually made a strenuous effort to understand it more thoroughly). While I did predict in a 1983 scanner survey that camera-based systems would be right along, I never dreamed that so much image processing and recognition software would be crammed into a KitKat-sized camera cell phone. I was skeptical that Wikipedia could become a useful and tolerably reliable source of information about pattern recognition. I hope that there are many more equally pleasant surprises coming down the pike!

One of my most agreeable duties from 1967 till 2007 was teaching a graduate course in pattern recognition. I always offered students completing the course with at least a B a lifetime guarantee to make myself available for any technical question that they might want to discuss. Some have taken me up on it. Now that I am retired, I have gone back to being a full time student. In addition to the occasional déjà vu, I look forward to learning much new material during SSSPR and ICPR 2012.

References

1 S. Veeramachaneni, P. Sarkar, and G. Nagy, "Modeling Context as Statistical Dependence," Proceedings of Modeling and Using Context: 5th International and Interdisciplinary Conference CONTEXT 2005, Paris, France, Springer Lecture Notes in Computer Science, vol.3554, pp. 515-528, July 5-8, 2005.

2 P.A. Devijver, “Relationship between statistical risk and the least-mean-square error criterion in pattern recognition,” Proceedings of the First International Joint Conference on Pattern Recognition, Washington, DC 1973.

3 PA Devijver, ”On a new class of bounds on Bayes risk in multihypothesis pattern recognition,” IEEE Trans. on, Computers, C-23,1, 70-80, 1974. 1974

4 P.A. Devijver and J. Kittler, Pattern recognition: A statistical approach, Prentice/Hall 1982

5 P.A. Devijver and C. Ronse, Connected Components in Binary Images: The Detection Problem John Wiley & Sons, Inc. New York, NY, USA ©1984

6 P.A. Devijver, “Classification in Markov Chains for minimum symbol error rate,” Proceedings of International Conference on Pattern Recognition, 1334-1336, 1984

7 Pierre A Devijver, “Baum's forward-backward algorithm revisited,” Pattern Recognition Letters, 3, 6, Dec. 1985, 369-373

8 P.A. Devijver and M. Dekesel, “Computing multidimensional Delaunay tessellations,” Patter Recognition Letters 4, 5-6, 311-316.

9 G. Nagy and X. Zhang, "Simple statistics for complex features spaces," Data Complexity in Pattern Recognition, pp. 173-195, M. Basu and T. K. Ho, Eds., Springer, 2006.

10 G. Nagy and G. L. Shelton, "Self-Corrective Character Recognition System," IEEE Trans. Information Theory, vol. 12, #2, pp. 215-222, April 1966.

11 G. Nagy, "The Application of Nonsupervised Learning to Character Recognition," Pattern Recognition, L. Kanal, Ed., pp. 391-398, Thompson Book Company, Washington, 1968.

12 H.S. Baird and G. Nagy, "A Self-correcting 100-font Classifier," Proceedings of SPIE Conference on Document Recognition, vol.SPIE-2181, pp. 106-115, San Jose, CA, February 1994.
13 Y. C. Ho and A. K. Agrawala, “On the self-learning scheme of Nagy and Shelton,” Proceedings of the IEEE, vol. 55, pp. 1764-1765, October 1967.
14 G. Nagy and N.G. Tuong, "On a Theoretical Pattern Recognition Model of Ho and Agrawala," Proceedings of the IEEE, vol. 56, #6, pp. 1108-1109, June 1968.
15 S. Veeramachaneni and G. Nagy, "Classifier adaptation with non-representative training data," Proceedings of the Fifth International Workshop on Document Analysis Systems, pp. 123-133, Princeton, NJ, Document Analysis Systems V, D. Lopresti, J. Hu, and R. Kashi, Eds., Springer LNCS 2423, August 2002.
16 R.G. Casey and G. Nagy, "Recognition of Printed Chinese Characters," IEEE Trans. Electronic Computers, vol. 15, #1, pp. 91-101, February 1966.
17 R.N. Ascher and G. Nagy, "A Means for Achieving a High Degree of Compaction on Scan-Digitized Printed Text," IEEE Trans. Computers, vol. 23, #11, pp. 1174-1179, October 1974.
18 R.G. Casey and G. Nagy, "Decision Tree Design Using a Probabilistic Model," IEEE Trans. Information Theory, vol. 30, #1, pp. 93-99, January 1984.
19 R.G. Casey and G. Nagy, "Autonomous Reading Machine," IEEE Trans. Computers, vol. 17, #5, pp. 492-503, May 1968.
20 R.G. Casey and G. Nagy, "Advances in Pattern Recognition," Scientific American, vol. 224, #4, pp. 56-71, 1971.
21 G. Nagy, S. Seth, and K. Einspahr, "Decoding Substitution Ciphers by means of Word Matching with Application to OCR," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 9, #5, pp. 710-715, September 1987
22 T.K. Ho and G. Nagy, "OCR with no shape training," Proceedings of International Conference on Pattern Recognition-XV, vol.4, pp. 27-30, Barcelona, Spain, September 2000.
23 G. Nagy. “Teaching a computer to read,” Proceedings of the Eleventh International Conference on Pattern Recognition, volume 2, pages 225–229, The Hague, 1992
24 S. Veeramachaneni and G. Nagy, “Analytical results on style-constrained Bayesian classification of pattern fields,” IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 29, #7, pp. 1280-1285, July 2007.
25 P. Sarkar and G. Nagy, "Style consistent classification of isogenous patterns," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, #1, pp. 88-98, January 2005.
26 S. Veeramachaneni and G. Nagy, "Style context with second order statistics," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, #1, pp. 14-22, January 2005
27 Y. A. Tijerino, D.W. Embley, Deryle W. Lonsdale, and G. Nagy, "Towards ontology generation from tables," World Wide Web Journal, vol. 6, #3, Springer- Verlag, 2005.
28 J. Zou and G. Nagy, “Visible models for interactive pattern recognition,” Pattern Recognition Letters Vol. 28, pp 2335-2342, 2007.