Invited plenary presentation, Procs. ICPR04, VOl. II, pp. 311-314, Cambridge UK, August 2004

Visual Pattern Recognition in the Years Ahead

George Nagy
DocLab, Rensselaer Polytechnic Institute, Troy, NY, USA
nagy@ecse.rpi.edu

Abstract

Conventional classification algorithms have already reached a plateau at the trade-off imposed by the bias due to the structure of the classifier and the variance due to the limited size of the training set. The latter may be alleviated by exploiting known constraints, including class and style priors, language models, statistical correlations between spatially proximate patterns, statistical dependence due to isogeny (common source) of patterns, and even information-theoretic properties of the representations that have evolved for symbolic patterns intended for communication. Another development that may lead to new applications of pattern recognition is more effective human intervention. The interplay of human and machine abilities requires models that are both human and computer accessible.

1. Beyond representative training sets

The fundamental assumption behind most trainable classifiers is that the test set is statistically representative of the training set. How representative is representative? The test set and the training set should be as similar as if they had been randomly chosen from a single corpus of patterns. In practice this assumption is often unworkable because drawing a random training sample requires prior access to the population to be classified. Even if enough representative training patterns were available, it would be too costly to label them.

We discuss two approaches to circumventing these limitations. The first is *constrained recognition*, and the second is explicit design for *optimal human intervention*. I will first discuss constrained recognition for high-volume tasks like OCR, where automatic classification is essential. Then I will discuss interactive recognition for low-throughput or sporadic applications, where accuracy is paramount. However, the two approaches are not incompatible: human intervention plays a part even in OCR. Note that the word "visual" in the title specifically excludes speech recognition and other signal processing applications, not because these ideas are inapplicable to them, but only to delimit the scope of discussion.

2. Constrained recognition

In addition to architectural differences among the data structures used to store and access decision boundaries, classifiers differ from one another by the type of information used to adjust their parameters. There has been much research, and significant recent progress, in "minimalist" pattern recognition, where the objective is to extract new information from sets of patterns represented by feature vectors (*unsupervised learning* [1,2]), from labeled vectors (*supervised learning* [3]), or from both labeled and unlabeled samples (*semi-supervised learning* [4]). The ultimate goal has been the development of *universal* classification algorithms based on arbitrary collections of feature vectors.

In contrast, the goal of *constrained recognition* is explicit and formal algorithmic application of whatever *domain-specific* information is available. The classifier should make use of all available contexts, much as a human would. Instead of recognizing patterns in isolation, it should recognize interrelated groups of patterns (*patterns of patterns*). Furthermore, we expect *dynamic classifier design* to foster learning during the entire operational lifetime of the classifier by making use of feedback from downstream use of its output [5]. Examples of pertinent constraints, starting with some that are already widely exploited, are listed below.

- The number and relative population of classes and subclasses (class and style priors) assume decisive importance with highly skewed demographics. Priors can be dynamically re-estimated according to strata defined by already extracted information, as in advanced form-processing systems [6].

- Statistical dependence among the *class labels* of adjacent patterns (linguistic context) has been modeled at the morphological, lexical, and syntactic levels [7,8], Many Hidden Markov Models make use of language models. The underlying transition probabilities need not be estimated from the same training set as the OCR features, because linguistic content is seldom correlated with typeface or handwriting.
- Statistical dependence among the *features* of adjacent patterns (due, for instance, to ligatures) occurs in handwriting (*allographs*), and in printed Arabic and Indic text.
- Class similarities may be predictable: the digit "1" is likely to be misrecognized as the letter "l". Confusion matrices are fairly stable over feature sets, typefaces, and writers. This constraint is occasionally applied to reduce alphanumeric confusions in forms reading.
- We have had some success in exploiting common-source constraints that give rise to statistical dependence among the features of (not necessarily adjacent) patterns in isogenous fields (*style context*) [9,10,11,12]. How Anne writes "*4*" may suggest how she writes "*9*". A more general notion of visual and structural style is presented in [13].
- Restrictions on the maximum difference between the class-means of test patterns and the class-means of the training samples lead to decision-directed approximation [14,15,16,17]. This constraint expresses the confidence that the classifier will recognize *most* of the patterns in the test set, even though they are statistically different from the training set. If so, the classifier can be re-trained with the newly labeled patterns. When does "unsupervised" adaptation work? While some sufficient conditions have been developed, they are much too restrictive for pattern distributions observed in practice. Necessary conditions have not been found.
- A common property of OCR feature spaces is that no linear combination of the mean vectors of a set of classes is located near the mean vector of any other class. Furthermore, the distribution of patterns about the class means is asymmetric: deviations away from the means of other classes (and from the mean of all patterns) tend to be larger than deviations towards other patterns. This affects outlier detection and may be the consequence of the evolution of scripts according to information-theoretic principles that preserve class-distinctions in the presence of noise [18]. It is less likely to hold for natural objects, such as multispectral crop signatures in remote sensing.
- The correlation between a pair of features depends much more on the chosen pair of features than on the class or style of the samples over which the correlation is computed. This constraint justifies various covariance matrix regularization schemes that are useful when there are not enough samples to accurately estimate class-covariances, but enough to estimate the overall covariance matrix.
- The order of the extracted features is preserved under elastic deformation of a sequence of patterns, as when handwriting is squeezed on approaching the right margin. Sequence-preserving inter-pattern relations are exploited in Symbolic Indirect Correlation (SIC). Nearest Neighbor classifiers are not directly applicable to unsegmented pattern sequences, but SIC is. Because such classifiers have high variance, they need *many* reference patterns. The development of algorithms that can perform the required comparison of bipartite graphs is only at a preliminary stage [19].

3. Interactive recognition

All operational pattern recognition systems make some use of human intervention. But except in Image Database, human interaction is seldom mentioned in research publications. We usually report only that the classifier was trained and tested on a "ground-truthed" database of so many patterns, and that 90% or 99% of the test patterns were correctly classified, without mentioning how much time was expended on producing the ground truth, or what is to be done with the 10% or 1% errors.

Even in operational systems, human intervention usually takes place only at the beginning and at the end. In OCR, for example, the scanned pages are usually inspected, and problematic pages are either set aside for manual data entry or rescanned with different settings. The segmentation of the document into text and non-text fields may also be inspected and corrected. At the other end of the pipeline, rejected words or fields are manually re-entered. In critical applications, either the entire output or a subset thereof (possibly selected by automated *triage* [20]) may be proofread in order to catch and correct OCR errors. Research on several other aspects of human interaction, listed below, would pay rich dividends.

- *Who is in charge?* It is tempting to think of the human as providing on-call assistance to a fallible automated classifier. It is more productive, however, to focus on how a computer can help human recognition. If human accuracy is desired, we cannot depend on the machine's assessment of when it is likely to be wrong. Confidence measures based on the tails of probability distributions are inherently untrustworthy. Therefore it is essential to let the human retain the initiative throughout the entire classification process. The role of the machine is only to save time by performing routine tasks under close supervision.

- How can we take advantage of the differences between human and machine cognitive abilities? Humans apply to recognition a rich set of contextual constraints and superior noise filtering abilities to excel in gestalt tasks, like object-background separation. Humans are also good at judging whether two images represent the same class. But computers can store thousands of images and associations between them, never forget a name or a label, and compute geometric moments and conditional probabilities. These differences suggest that a system that combines human and machine abilities can, in some situations, outperform both. (And have also inspired research on the design of tasks that unfailingly favor humans [21]!)
- The paramount design criterion for interactive pattern classification must be the *minimization of human labor for a given level of recognition accuracy*. Since the accuracy of a well-designed interactive system is governed by human accuracy, the system needs to be fast enough only relative to human reaction time, which can be measured in billions of machine cycles. Interaction is profitable where higher accuracy is required than is currently achievable by automated systems, but when there is enough time for limited human interaction. In such problem domains, the fundamental research question is when, where, and how to interact [22].
- *Direct interaction with images* was demonstrated recently in the narrow domains of face and sign recognition. However, it was confined to pre-processing, i.e., establishing the pupil-to-pupil baseline [23] or text bounding-box [24,25]. Effective interaction is required throughout the classification process, rather than just at the beginning and the end. It appears that the most appropriate channel for interaction is a *parameterized domain-specific visual model* [26]. Such a model is an abstraction of the overall shape of an object and of its most discriminative constituents. For a face, it may be the contours of the head, chin, temples, eyes and nose. For flowers, it could the petals and leaves. Such a model is meaningful to both the human operator and to the computerized feature extraction system. Fitting it to an unknown object requires only weak segmentation, if necessary assisted by human gestalt perception, rather than familiarity with the distinguishing features of the classes. Either man or machine can accomplish it, with varying degrees of success. We have not, however, found any previous work advocating *iterative* image-based interaction to bridge the "semantic gap [27]."
- *Content-based image retrieval vs. object classification.* The evaluation criteria for image retrieval are usually *precision* and *relevance*, while for objective classification they are *accuracy* or *recognition rate*. The most advantageous methods of interaction are not necessarily the same in broad domains, like a personal photo collection or a tourist web site, as in narrow domains, like flower or face recognition. In the broad domains of content-based image retrieval, relevance feedback has been found effective [28]. Interaction is, however, necessarily limited to the choice between acceptable and unacceptable responses, because no effective way has been found so far to interact with arbitrary images without a domain-specific model. In object recognition, as opposed to exploratory data analysis [29], direct interaction with features is unlikely to pan out because our conception of the disposition of patterns in a high-dimensional space is likely to be poor.
- *Model-based feature extraction.* To rank-order candidate classes, features should be extracted algorithmically from the image according to the current model of the object, which may be as simple as a coarse approximation of its boundary. The user should be able to interact with the image anytime that he or she considers the computer's response unsatisfactory. The interaction extracts some features (i.e., discriminatory model parameters) directly, and improves the accuracy of other extracted features indirectly, by improving the fit of the computer-proposed model. Eventually the user selects the appropriate candidate from the ordered reference images.
- *Learning from interaction.* The computer must make subsequent use of the parameters of the improved models to improve not only its own statistical model-fitting process, but also its internal classifier. Classifier adaptation can be based on human confirmation of the final identification, which is likely to be almost error-free. The automated parts of the system will gradually improve, and decrease the need for human intervention. As an important byproduct, the operator's judgment of when interaction is beneficial will also improve. Experiments on flower and face recognition demonstrate both phenomena [30].
- *Mobile systems.* Personal digital assistants (PDAs) with touch-sensitive screens and even camera phones now have enough processing power and storage to support interactive recognition. They offer the enormous advantage of being able to take additional pictures of an object during the classification process.
- *Open Mind Initiative.* Interactive recognition systems networked to each other can benefit from distributed data collection and interaction by multiple operators. Mobile systems could operate in a wireless client-server mode, with the classification performed on a server with access to many such systems.

4. Conclusion

We expect, with perennial optimism, that in the years ahead vastly enlarged classification contexts, and judicious application of close-coupled human intervention, will lead to significant expansion in the application of visual pattern recognition systems. These techniques are synergistic: each improves the benefit from the other. My talk will highlight relevant examples drawn from OCR, forms processing, botany, physiognomy, and dermatology.

References

[1] A.K. Jain, P. Flynn, Data clustering: A review, *ACM Computing Surveys 31*(3): 264-323, 1999.

[2] M. Figueiredo, A.K. Jain, Unsupervised learning of finite mixture models, *IEEE Trans. on PAMI 24*: 381-396, 2002.

[3] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: A review, *IEEE Trans. on PAMI 22*(1):4-37, 2000.

[4] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, *Procs. Workshop on Computational Learning Theory (COLT)*, Morgan Kaufmann, 1998.

[5] G. Nagy, Classifiers that improve with use, *Procs. Conference on Pattern Recognition and Multimedia), IEICE*, Tokyo, 79-86, February 2004.

[6] B. Klein, A.R. Dengel, Problem-adaptable document analysis and understanding for high-volume applications, *IJDAR 6* (3), 167-180, March 2004.

[7] H.S. Baird, K. Thompson, Reading Chess, *IEEE Trans. on PAMI 12* (6), 552-559, June 1990.

[8] T.K. Ho, G. Nagy, OCR with no shape training, *Procs. ICPR XV, Vol 4*, Barcelona, 27-30, Sept. 2000.

[9] P. Sarkar and G. Nagy, Style-consistency in isogenous patterns, *Procs. 6th Int'l Conference on Document Analysis and Recognition*, Seattle, 1169-1174, Sept. 2001.

[10] P. Sarkar, An iterative algorithm for optimal style-conscious field classification, *Procs. ICPR XVI*, Vol. 4, 243-246, Quebec City, Canada, August 2002.

[11] S. Veeramachaneni, G. Nagy, C.-L. Liu, and H. Fujisawa, Classifying isogenous fields, *Proc. 8th Int'l Workshop on Frontiers of Handwriting Recognition* (IWFHR02), Niagara-on-the-Lake, Canada, 41-46, August 2002.

[12] S. Veeramachaneni and G. Nagy, Adaptive classifiers for multisource OCR, *Int'l J. Document Analysis and Recognition 6*(3), 154-166, March 2004.

[13] A.D. Bagdanov, Style characterization of machine printed text, PhD thesis, University of Amsterdam, 2004.

[14] G. Nagy and G. L. Shelton Jr., Self-corrective character recognition system, *IEEE Trans. Information Theory*, IT-12(2): 215-222, April 1966.

[15] G. Nagy, The application of nonsupervised learning to character recognition, in L.N. Kanal (Editor), Pattern Recognition, Thompson , Washington, 391-398, 1968.

[16] H. S. Baird and G. Nagy, A self-correcting 100-font classifier, In L. Vincent and T. Pavlidis (editors), *Document Recognition*, SPIE Vol.2181, 106–115, 1994.

[17] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Wiley, 2001.

[18] G. Nagy and S. Veeramachaneni, A Ptolemaic model for OCR, *Proc. 7th Int'l Conference on Document Analysis and Recognition*, Edinburgh, 1060-1064, August 2003.

[19] G. Nagy, A. Joshi, M. Krishnamoorthy, D.P. Lopresti, S. Mehta, S. Seth, A nonparametric classifier for unsegmented text, *Procs. Document Recognition and Retrieval XI*, SPIE Vol. 5296, 102-108, January 2004.

[20] P. Sarkar, H.S. Baird, J. Henderson, Triage of OCR output using 'confidence" scores, in *Procs. Document Recognition and Retrieval IX*, SPIE/IS&T, San Jose, 2002.

[21] H.S. Baird, A.L. Coates, R.J. Fateman: Pessimal print: a reverse Turing test. *Int'l J. Document Analysis and Recognition 5*(2-3):158-163 (2003).

[22] G. Nagy, J. Zou, Interactive visual pattern recognition, *Procs. ICPR XVI*, Vol. III, Quebec City, 478-481, Aug. 2002.

[23] J. Yang, X. Chen, and W. Kunz, A PDA-based Face Recognition System, *Proc. of the 6th IEEE Workshop on Applications of Computer Vision*, 19-23, December, 2002.

[24] J. Zhang, X. Chen, J. Yang, and A. Waibel, A PDA-based Sign Translator, *Proc. of the 4th IEEE Int. Conf. on Multi-modal Interfaces*, 217-222, 2002.

[25] I. Haritaoglu, "Scene Text Extraction and Translation for Handheld Devices," *Proc. CVPR01*, vol. 2, pp. 408-413, 2001.

[26] J. Zou, G. Nagy, Evaluation of model-based interactive pattern recognition, *Procs. ICPR XVII*, August 2004, in press.

[27] A.W. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, *IEEE Trans. on PAMI 22*(12): 1349-1380, 2000.

[28] Y. Rui, T.S. Huang, M. Ortega, and S. Mehrotra, Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval, *IEEE Trans. Circuits and Systems for Video Technology 8*(5), 5644-655, 1998.

[29] T.K. Ho, Exploratory Analysis of Point Proximity in Subspaces, *Procs. ICPR XVI, Vol. 2, 196-193*, Quebec City, August 2002.

[30] J. Zou, Computer Assisted Visual InterActive Recognition – CAVIAR, PhD thesis, Rensselaer Polytechnic Institute, Troy, NY, May 2004.