Advanced Character Recognition 6610
(invited)

George Nagy
Rensselaer Polytechnic Institute, Troy, NY, USA

CATALOG DESCRIPTION

ECSE 6610 - Advanced Character Recognition. ***Principles and practice of the recognition of isolated or connected typeset, hand-printed, and cursive characters. Review of optical digitization, supervised and unsupervised estimation of classifier parameters, bias and variance, expectation maximization, the curse of dimensionality. Advanced classification techniques including classifier combinations, support vector machines, hidden Markov methods, styles, language context, adaptation, segmentation-free classifiers, indirect symbolic correlation.*** **Prereq*: ECSE 2610, Probability, Linear Algebra.* Spring term annually.**

ECSE-6610 FIRST DAY HANDOUT

Instructor:	Prof. George Nagy
Office hours:	After class in the bar
Email:	nagy@ecse.rpi.edu

Text: S. V. Rice, G. Nagy, T. A. Nartker
Optical Character Recognition:
An Illustrated Guide to the Frontier [RNN 99]

Reference texts (on reserve at Folsom Libe):

Duda, Hart, & Stork, Wiley 2001	[DHS 01]
Mitchell, McGraw-Hill 1997	[MT 97]
Nadler & Smith, Wiley 1993	[NS 93]
Schürmann, Wiley 1996	[SJ 96]
Theodoridis & Koutroumbas, Acad.1999	[TK 99]
Vapnik, Wiley 1998	[VV 98]

For additional sources, see the Text and the Bibliography.

Grading: Five programming assignments
Term Paper
Final Examination

SYLLABUS

1. Review: Intro to OCR (ECSE 2610)

Preprocessing:
Scanner calibration, correction of scan distortions; noise removal; text-figure separation; skew correction; gray-scale and color, text layout extraction (column, line, and word segmentation) [NG 00].
Character image defect models [KBH 94]
Recovery of scanner distortions [BE 00].
Help Session: Wed. 6 pm Prof. E. Barney Smith.

Features:
Reflectance, geometric, & topological invariants [FG 60, SM 61]
Features as weak classifiers [KE 00]
N-tuples and feature selectection
[JN 95, JDM 00, JKNS 96]

Resource person: Dr. D-M Jung, Yahoo!

Static single-pattern classifiers:
Bayes: Single & Multimodal, Linear, Quadratic, Gaussian and Bilevel [DHS 01, LKF 01]
Neural Networks: Backprop, LVQ, RBF [BC 95]
Support Vector Machine [VV 98]
Nearest Neighbors [DHS 01]
Decision Trees and Forests [AGW 07, TH 98]

Classifer training:
Sample size and dimensionality [RJ 91]
Bias and variance [GBD 92]
Bagging, Boosting, Random Subspaces [JDM 00]
Clustering [TK 99]
Expectation Maximization [DLR 77, RW 84]

Generalization, validation, and error prediction:
Validation, Jackknife, Bootstrap
[DHS 01, JDM 00].

2. Adaptive classifiers

Unlabeled samples may suffice to identify the parameters of separable distributions. Then a small number of labeled samples can be used to determine which distribution corresponds to which class [CC 95, 96]. One method to accomplish this is to use the labels assigned by a relatively accurate classifier to generate larger, more representative training sets. This may be useful when the test set is similar to a subset of the training set (i.e., specialization of an omnifont classifier to a single font) [NS 66, BN 94].

3. Classifier combinations

The motivation for classifier combination is that different classifiers, feature sets, different training sets, and different regions of the feature space may provide complimentary information. *Decision optimization* attempts to combine, in the best possible way, a set of complete, fully trained classifiers. *Coverage optimization* aims to tune, with varying degrees of success, each classifier to different aspects of the training set [HT 01]. A possible means of achieving this is offered by the theory of Stochastic Discrimination.

Stochastic Discrimination proposes to combine many weak classifiers that are just subsets (*models*) of the feature space. As more models are added, the combination converges to optimally separate the classes, provided that three conditions are met: *(1) Discrimination* or *enrichment* (each constituent must favor some class)*, (2) Uniform Coverage* (the ensemble of models does not favor any populated region over any other) and *(3) Indiscernability* (the characteristics of the sample space of weak models with respect to a point in the feature space must be the same whether that point is a training point or a test point). The major advantage claimed is immunity (or at least resistance) to over-training with increasing classifier complexity [KE 00].

The theory has been applied to constructing decision forests of weak classifiers based on random splits of the feature space. Each classifier for a specific subspace is invariant for points that differ from the training samples *only* in the orthogonal (unselected) subspace, therefore such a classifier can improve its generalization accuracy even as it grows in complexity [HT 98].

Resource person: Dr. Tin Kam Ho (Bell Labs)

4. Styles: source-dependent distributions

In many OCR tasks, patterns appear in groups (*fields*) that have common traits (*style*) because of their common origin (writer, font, printer, scanner). In addition to *rendering style*, one may also consider *linguistic style* (favored letter or word sequences), and combinations of the two. If some of the recognition errors are due to inter-style rather than intra-style confusions, then the statistical dependence between features of different characters within the same field can be exploited to improve recognition accuracy.

Optimal style-conscious recognition differs from font or writer recognition because the probabilities of the constituent styles are taken into account instead of a winner-take all approach. It also differs from the use of context, because we model the dependence of features in a field, rather than the interdependence of their class labels.

Several sources may share the same style with respect to any class. The estimation of style-conscious classifier parameters requires a style-unsupervised approach (e.g., Expectation Maximization) if, as is usually the case, the styles are not specifically labeled in the training set [SN 99, SN 00, SN 01].

Help Session: Wed. 6 pm, Dr. Prateek Sarkar

Small-sample style estimation for quadratic classifiers. Linear and quadratic discriminant methods have been shown time and again to provide robust yet versatile classification. In principle a quadratic field classifier can simply be trained on field samples to exploit style. However, with 100 features, 20 classes, and a field length of only four, a single sample of every field class requires 160,000 four-hundred-dimensional vectors, as opposed to 20 one-hundred-dimensional vectors for a singlet classifier.

To avoid this exponential growth in the number of required training samples, assume that the class-conditional variability of the patterns classes is independent from pattern to pattern within a field from the same source. (The features of a singlet pattern retain their dependence.) In this case, the correlation of the feature vectors within a field can be estimated from the *covariance matrix of the mean vectors of each field* and the (smaller) singlet class-conditional feature covariance matrices.

Although the above assumption mitigates the Curse of Dimensionality, we have not yet found a way to avoid the linear growth in the covariance matrices at classification time. We therefore consider the application of the method only to salvaging rejects.

Help Session: Wed. 6 pm, H. Veeramachaneni

5. Linguistic Context

Isolated ambiguous patterns can often be identified from context if their neighbors are known. What if none of the identities are known? We are then faced with a cipher-substitution problem that can be solved by generating a mapping that satisfies certain linguistic criteria, from groups of equi-shaped patterns to the alphabet. Difficulties arise when the patterns that correspond to the same alphabetic symbol are clustered into several groups (as in the case of upper and lower case versions of a letter), and when the unknown text has a low ratio of sample size to the number of classes. Lexicons can be used for unusual typefaces [HN 00], and symbol-n-grams for distinguishing cases, digits, and punctuation [HN 01]. Word context is used in postal address readers [NG 92, LSF 01].

6. Segmentation-free classifiers

In the production of training samples for typeset text, manual character segmentation takes far more time than transcription. Character widths, however, can be estimated from word-segmented samples and their transcripts. Isolated prototypes for each character can then be extracted automatically by estimating the approximate character positions using the cumulative character widths, and finding the peak correlation between the bitmaps of words which share one or more letters [XN 99].

Hidden Markov Methods represent Markov models where the observable features of each state are generated from a probability distribution, so that the state transitions cannot be estimated directly. Efficient algorithms are available, but convergence depends on appropriate initialization. In OCR, the *grain* of the model is often finer than the size of the characters, with each character represented by a separate model. A coarse-grained model may also be used to represent character sequences [RL 89].

Communication-Channel Models unite some aspects of preprocessing (e.g., baseline finding), layout analysis, and character recognition. A block of text is modeled as a Markov source whose transitions generate character placements, white spaces, line feeds, carriage returns, and character bitmaps that are degraded by printing and scanning ("channel noise"). The decoding process, based on dynamic programming, attempts to identify the most likely sequence of transitions from the observed pixels. The required input consists of character bitmaps, transition probabilities of character classes and layout-generating operations. Efficient algorithms have been found to bootstrap the estimators from very limited training samples [KL 97].

Resource person: Dr. Yihong Xu, EMC.

N-gram based classifiers for unsegmented words. Anagrams, or different words composed of the same letters, are fairly common in English. However, words that share exactly the same letter-bigrams, are rare. To recognize a word without character-level segmentation, we can correlate it with a set of words with known identities (called *reference patterns*). If the extent of correlation suffices to determine whether two words share a letter bigram, then we can devise elegant algorithms to recognize words from a large lexicon, using only a limited number of identified (but *unsegmented*) reference words. Recognized words can be added to the reference string without further processing to increase recognition accuracy [EN 00, EVN 01].

Help Session: Wednesday 6 pm,
Adnan El Nasan & Harsha Veeramachaneni

7. Indirect Symbolic Correlation ***new***

The above bigram-based method does not use *all* of the order information in the text. The following approach was devised specifically for individualized large-vocabulary human-computer communication via stylus or audio input. It represents a major departure from current paradigms.

The recognition is based on two levels of sequence comparisons. At the first level, each word in a lexicon of *n* allowable words is compared to a transcript of a reference subset of the known words or phrases, resulting in *n lexical match graphs*. The arbitrary feature-string representation of an unknown word or phrase is compared to the feature-string representation of the reference words (ordered as in the lexical comparison) to identify similar segments. This yields the *feature match graph.*

At the second level, the feature match graph is compared with each lexical match graph. The similarity of two graphs is expressed in terms of the maximum clique of their *compatibility graph*. Efficient heuristic algorithms are available for computing the maximum clique.

The unknown word must be part of the lexicon but, in contrast with whole-word recognition, need not be part of the reference set. In contrast to Hidden Markov Methods, no probabilities need to be estimated. The probability that a similar ordering of the feature-level and lexical matches identifies the correct word increases rapidly with the length of the reference signal, and decreases slowly with the size of the lexicon of admissible words [NH 00].

SNAP TEST

1. Does Bagging reduce classifier bias, variance, or neither? Justify your answer.
2. Can feature correlation reduce classification error over uncorrelated features with the same class-conditional means?
3. If two separate sets of features are available, is it better to combine feature information before or after classification? Why?
4. What are the three essential conditions for Stochastic Discrimination?
5. What is the penalty incurred by style-conscious classification over singlet classification?
6. Where and why is EM used in HMM classification?
7. How can unlabeled samples improve classifier accuracy? Give an example where the classification after adaptation is worse.
8. Find two English words of at least seven letters that share exactly the same bigrams.
9. What part do Maximal Cliques play in Indirect Symbolic Correlation?

BIBLIOGRPAPHY

[AGW 97] Y. Amit, D. Geman, K. Wilder, Joint induction of shape features and tree classifiers, *IEEE-PAMI 19*, 11, 1300-1305, 1997.

[BC 95] C.M. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, 1995.

[BE 01] E. Barney Smith, Scanner Parameter Estimation Using Bilevel Scans of Star Charts, *Proc. ICDAR 2001.*

[BN 94] H.S. Baird & G. Nagy, A self-correcting 100-font classifier, *Proc. DR&R, SPIE-2181,* 106-115, San Jose, February 1994.

[CC 95] V. Castelli & T.M. Cover, On the exponential value of labeled samples, *PRL 16*, 105-111, 1995.

[CC 96] V. Castelli & T.M. Cover, The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter, *IEEE-IT 42*, 6, 2101-2117, 1996.

[DHS 01] R.O. Duda, P.E. Hart, & D.G. Stork, Pattern Classification, Wiley, 2001.

[DLR 77] A.P. Dempster, N.M. Laird, & D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Royal Statistical Society Series B, 39,* 1-38, 1977.

[EN 00] A. El-Nasan & G. Nagy, InkLink, *Proc. ICPR-XV,* Vol. 2, 573-576, Barcelona, 2000.

[EVN 01] A. El-Nasan, H. Veeramachaneni, G. Nagy, Word discrimination based on Bigram Co-occurrences, *Proc. ICDAR 2001.*

[FG 60] G.L. Fisher (ed.), Optical Character Recognition, Spartan Books, Washington, 1960

[GBD] S. Geman, E. Bienstock, R. Doursat, Neural networks and the bias/variance dilemma, *Neural Computation 4*, 1, 1-58, 1992.

[HN 00] T.K. Ho & G. Nagy, OCR with no shape training, *Procs. ICPR-XV,* Vol. 4, 27-30, Barcelona, September 2000.

[HN 01] T.K. Ho & G. Nagy, Exploration of contextual constraints for character pre-classification, *Proc. ICDAR 2001.*

[HT 98] T.K. Ho, The random subspace method for constructing decision forests, *IEEE-PAMI-20*, 8, 832-844, August 1998.

[HT 01] T.K. Ho, Multiple classifier combination: lessons and next steps, in Hybrid Methods in Pattern Recognition (H. Bunke, A. Kandel, editors), World Scientific 2001.

[JKNS 96] D.M. Jung, M.S. Krishnamoorthy, G. Nagy, A. Shapira, N-tuple features for OCR revisted, *IEEE-PAMI 18,* 7, July 1996.

[JDM 00] A.K. Jain, R.P.W. Duin, & J. Mao, Statistical pattern recognition: A review, *IEEE-PAMI-22,* 1, 4-37, January 2000.

[JN 95] D.M. Jung & G. Nagy, Joint classification and feature extraction for OCR, *Proc.* ICDAR-95, Montreal, 1115-1118, 1995.

[KBH 94] T. Kanungo, H.S. Baird, & R.M. Haralick, Validation and estimation of document image degradation models, Procs. *4th ISRI Symposium on Document Analysis and Retrieval,* 217-228, 1994.

[KE 00] E.M. Kleinberg, On the algorithmic implementation of stochastic discrimination, *PAMI-22*, 5, May 2000, 67-76.

[KL 97] G.E. Kopec, & M. Lomelin, Supervised template estimation for document image decoding *PAMI-19,* 12, 1313-1324, 1997.

[LKF 01] C.L. Liu, M. Koga, H. Fujisawa, Lexicon-Driven Handwritten Character String Recognition for Japanese Address Reading, *ICDAR-01.*

[LSF 01] C.L. Liu, H. Sako, H. Fujisawa, Performance evaluation of pattern classifiers for handwritten character recognition, *IJDAR*, in press 2001.

[MT 97] T.M. Mitchell, Machine Learning, McGraw-Hill 1997.

[NG 00] G. Nagy, Twenty years of document image analysis in PAMI, *IEEE-PAMI-22*, 1, 38-62, January 2000.

[NG 92] G. Nagy, Teaching a computer to read (invited), *Proc. ICPR-11, Vol. 2,* pp. 225-229, The Hague, August 1992.

[NH 00] G. Nagy & H. Veeramachaneni, personal communication, Troy, NY, 2000.

[NS 66] G. Nagy, & G. Shelton, Self-Corrective Character Recognition System, *IEEE IT-12,* #2, 215-222, April 1966.

[NS 93] M. Nadler & E.J. Smith, Pattern Recognition Engineering, Wiley, New York 1993.

[RJ 91] S.J. Raudys, & A.K. Jain, Small sample effects in statistical pattern recognition: Recommendations for practitioners, *IEEE-PAMI 13*,3, pp. 252-263, 1991.

[RL 89] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE 77* (2), 1989.

[RNN 99] S. Rice, G. Nagy, T. Nartker, Optical Character Recognition: An Illustrated Guide to the Frontier, Kluwer 1999.

[RW 84] R.A. Redner & H.F. Walker, Mixture densities, maximum likelihood, and the EM algorithm, *SIAM Review 26*, 2, pp. 195-235, 1984.

[SJ 96] J. Schürmann, Pattern Classification, Wiley, New York, 1996.

[SM 61] M.E. Stevens, Automatic character recognition - State-of-the-art report, National Bureau of Standards & Technology, Tech. Note 112, Washington 1961.

[SN 99] P. Sarkar & G. Nagy, Heeding more than the top template, *Proc. ICDAR 2001,* Bangalore, 1999.

[SN 00] P. Sarkar & G. Nagy, Classification of style-constrained pattern fields, *Procs. ICPR-XV,* Vol. 2, pp. 859-862, Barcelona, Sept. 2000.

[SN 01] P. Sarkar & G. Nagy, Style-consistency in isogenous patterns, *Proc. ICDAR 2001.*

[TK 99] S. Theodoridis & K. Koutrumbas, Pattern Recognition, Academic Press, 1999.

[VV 98] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, 1998.

[XN 99] Y. Xu & G. Nagy, Prototype Extraction and Adaptive OCR, *IEEE-PAMI-21*, 12, pp. 1280-1296, Dec. 1999.