

## A NONPARAMETRIC CLASSIFIER FOR UNSEGMENTED TEXT

George Nagy                      Rensselaer Polytechnic Institute  
 Ashutosh Joshi                  Rensselaer Polytechnic Institute  
 Mukkai Krishnamoorthy       Rensselaer Polytechnic Institute  
 Yu Lin                              University of Nebraska - Lincoln  
 Dan Lopresti                      Lehigh University  
 Shashank Mehta                  IIT Kanpur  
 Sharad Seth                       University of Nebraska - Lincoln

3/16/04

DR&amp;R 2004 Nagy et al.

1

## SYMBOLIC INDIRECT CORRELATION (SIC)

**SYMBOLIC:** it matches ordered *symbol* polygrams

**INDIRECT:** it is based on a *comparison of comparisons*

**CORRELATION:** uses sliding windows

It is a new idea, so far untested on real data.

Recognition based on *local matches* between unsegmented patterns at both the feature level and the lexical level

Testing on print, on-line handwriting, and speech in progress.

3/16/04

DR&amp;R 2004 Nagy et al.

2

## ADVANTAGES

- Matches based on signal subsequences of any length, although they are typically longer than single characters or phonemes (e.g. polygrams).
- Accommodates common distortions in camera- and tablet-based OCR (stretching, contraction) and speech (time-warping).
- Does not depend on the specific medium, feature set, and vocabulary.
- No training, only a reference set as in Nearest Neighbors, therefore suitable for unsupervised adaptation.
- Extensible to phrase recognition.

3/16/04

DR&amp;R 2004 Nagy et al.

3

## APPROACH

1. **Lexical Matching.** Match *polygrams* in every lexicon word against the transcription of the reference-signal (preprocessing).
2. **Feature Matching.** Match feature strings derived from the query and reference signals.
3. **Graph Matching.** Match the feature graph (Step 2) against the lexical graph (Step 3) for each word in the lexicon.
4. **Result.** Output the best matching lexicon word in step 3 as the result.

3/16/04

DR&amp;R 2004 Nagy et al.

4

## SIC example

### Reference string:

period                      ever                      people

3421213123332412124321~21341314213123~342121334211122213

### Unknown words (in lexicon)

*lever* :                      112221341314213123  
*perplex* :                      3421213123342111222131334

### Location of lexical matches in the reference string:

*lever*

*lever* *period*  
*lever* *ever*  
*lever* *people*

*perplex*

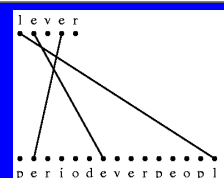
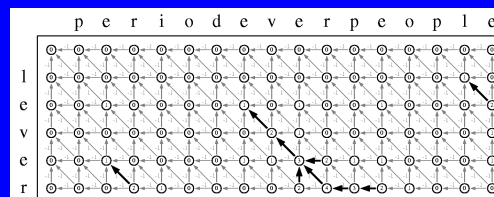
*perplex* *period*  
*perplex* *ever*  
*perplex* *people*

3/16/04

DR&amp;R 2004 Nagy et al.

5

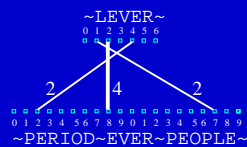
## LEXICAL STRING MATCHING



3/16/04

6

## Lexical match graph



(Bigrams and higher polygrams)

3/16/04

DR&R 2004 Nagy et al.

7

## Approach

1. **Lexical Matching.** Match *polygrams* in every lexicon word against the transcription of the reference-signal set (preprocessing).
2. **Feature Matching.** Match feature strings derived from the query and reference signals.
3. **Graph Matching.** Match the feature graph (Step 2) against the lexical graph (Step 3) for each word in the lexicon.
4. **Result.** Output the best matching lexicon word in step 3 as the result.

3/16/04

DR&R 2004 Nagy et al.

8

## Feature Extraction

(Courtesy: Adnan El-Nasan)

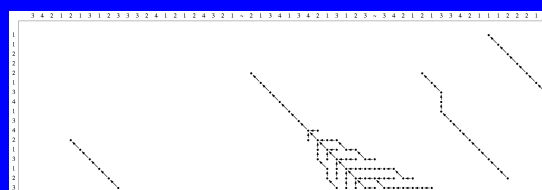


3/16/04

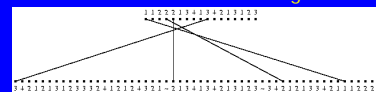
DR&R 2004 Nagy et al.

9

## FEATURE STRING MATCHING



Unknown feature string



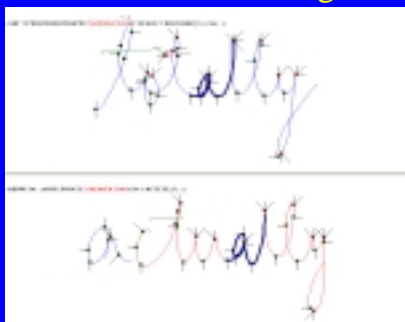
Reference feature string

3/16/04

DR&R 2004 Nagy et al.

10

## Feature Matching



3/16/04

DR&R 2004 Nagy et al.

11

## Approach

1. **Lexical Matching.** Match *polygrams* in every lexicon word against the transcription of the reference-signal set (preprocessing).
2. **Feature Matching.** Match feature strings derived from the query and reference signals.
3. **Graph Matching.** Match the feature graph (Step 2) against the lexical graph (Step 3) for each word in the lexicon.
4. **Result.** Output the best matching lexicon word in step 3 as the result.

3/16/04

DR&R 2004 Nagy et al.

12

### Graph Matching: hypothesis = "perplex"

~LEVER~  
0 1 2 3 4 5 6 7 8 9 0 1 2

FEATURE GRAPH

~PERIOD~EVER~PEOPLE~  
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 8 9 0 1 2 3 4 5 6 7 8 9 0

LEXICAL GRAPH

~PERPLEX~  
0 1 2 3 4 5 6 7 8

Order Isomorphic subgraphs

- Best matching subgraphs
- Preserve Edge Crossings, i.e. find *order isomorphic subgraphs*

3/16/04 DR&R 2004 Nagy et al. 13

### Graph Matching-2 hypothesis = "lever"

~LEVER~  
0 1 2 3 4 5 6 7 8 9 0 1 2

FEATURE GRAPH

~PERIOD~EVER~PEOPLE~  
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 8 9 0 1 2 3 4 5 6 7 8 9 0

LEXICAL GRAPH

~LEVER~  
0 1 2 3 4 5 6

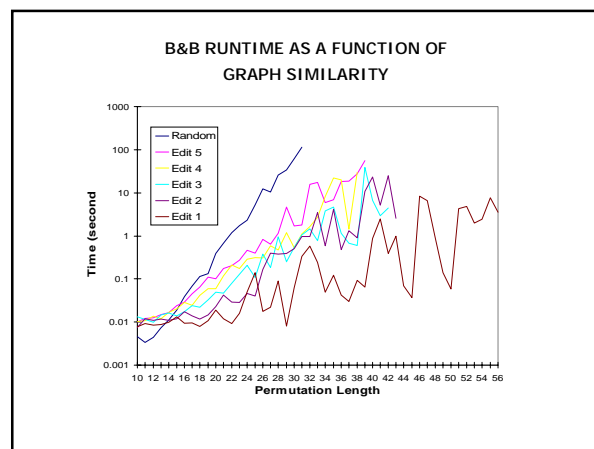
3/16/04 DR&R 2004 Nagy et al. 14

### FORMAL PROBLEM STATEMENT

$$L^* = \underset{i}{\operatorname{argmax}} \{ C_M( M_V(V_x, V_{Ref}), M_L(L_i, L_{Ref}) ) \}$$

$V$  = feature string  
 $L$  = lexical string  
 $M$  = string comparison (via *Smith-Waterman* algorithm)  
 $C_M$  = meta-comparison operator on two bipartite graphs  
 - a new *branch-and-bound* algorithm that finds the longest common subsequence (permutation).

3/16/04 DR&R 2004 Nagy et al. 15



### Simulation experiments

- Selection of
  - Lexicon and
  - Reference String
- Noise Models
- Implementations and Results

3/16/04 DR&R 2004 Nagy et al. 17

### Reference Set

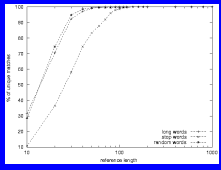
1000 words Selected in three different ways from the Brown Corpus:

1. common, short words (*stop words*)
2. long uncommon words
3. random

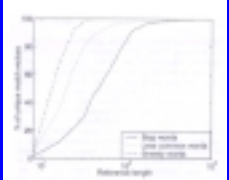
3/16/04 DR&R 2004 Nagy et al. 18

### Lower bounds on error rate

- Determine how the percentage of unique (correct) matches grows with the reference-string size and compare with the same for bigram co-occurrence.



Graph Matching



Bigram Co-occurrence

[El-Nasiri, Veermachani, Nagy, ICDAR 2001]

3/16/04
DR&R 2004 Nagy et al.
19

### Noise Model

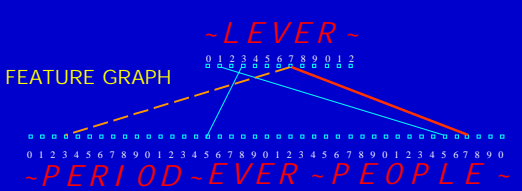
- Noise Level: normalized parameter  $O$  in  $[0,1]$  range.
- Symmetric Noise Model:
 
$$O = p(e|0) + p(e|1) \text{ and } p(e|0) = p(e|1)$$
- Weighted Noise Model:
 
$$O = w1 * p(e|1) + (1-w1) * p(e|0)$$
 where  $w1$  = size of query graph normalized wrt the size of the complete bipartite graph

3/16/04
DR&R 2004 Nagy et al.
20

### Feature Matching Errors

**-LEVER-**

0 1 2 3 4 5 6 7 8 9 0 1 2



0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0

**-PERIOD-EVER-PEOPLE-**

Error Types:

- Extra Edge (False Positive Error)
- - - - Missing Edge (False Negative Error)

3/16/04
DR&R 2004 Nagy et al.
21

### Implementation

- Graph size check:** Eliminate lexical graphs that differ substantially in size from the query graph. (Only surviving lexical words used in subsequent matching).
- Reference String Partitioning:** 1000-word reference string partitioned into 100 substrings of 10 words each.
- Matching:** Query word matched with lexical words independently for each substring and match scores accumulated.
- Recognition Result:** Top-scoring lexical word

3/16/04
DR&R 2004 Nagy et al.
22

### Simulation results

#### Symmetric Noise Model

Q (noise):	0.1	0.2	0.3	0.4	0.6	0.8	1.0
% Rec. Rate:	99.2	99.1	98.7	98.3	96.0	85.5	66.6

#### Weighted Noise Model

Q:	0.4	0.6	0.8	1.0
% Rec. Rate:	100	100	100	96.0
Reference String Size:	390	500	600	750
$Mean \left( \frac{Top^p - Top^q}{Top^q} \right)$	0.20	0.16	0.14	0.13

3/16/04
DR&R 2004 Nagy et al.
23

### ANOTHER IMPLEMENTATION WITH DIFFERENT HEURISTICS

Q	min ref	max ref	med ref	correct	wrong	rejected
0.1	10	740	60	95.8%	1.5%	2.7%
0.2	10	620	70	87.1%	7.7%	5.2%

Q	min ref	max ref	med ref	correct	wrong	rejected
0.1	10	620	40	96.9%	1.9%	1.2%
0.2	10	680	40	83.8%	8.7%	7.5%

3/16/04
DR&R 2004 Nagy et al.
24

## EXPERIMENTS ON (TOY) PRINTED CHARACTERS

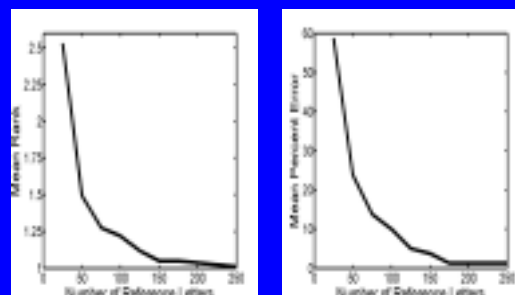


3/16/04

DR&amp;R 2004 Nagy et al.

25

## RANK AND ERROR RATE ON PRINTED CHARACTERS



3/16/04

DR&amp;R 2004 Nagy et al.

26

## RESEARCH QUESTIONS

- Is the Branch & Bound algorithm efficient enough or should we examine approximate clique algorithms?
- Does the error rate converge to the NN-Bayes bound?
- What is the best strategy for selecting reference words?
- Extension to phrases?
- What is the best cost function and algorithm for feature matching?

3/16/04

DR&amp;R 2004 Nagy et al.

27

~~CONCLUSIONS CLAIMS~~

- No parameter estimation required, therefore easy to update. Recognition should improve with use. Humans adapt faster than machines but static systems cannot exploit user learning.
- Inherently more accurate than character-based scheme. Lexical context is essential for speech and handwriting recognition.
- New words can be added to lexicon without changing the ref list. Many applications require a large, easily extensible vocabulary.
- Current training methods, based on HMMs are unfriendly and unrepresentative.
- Source-specific recognition is easier:  
in most human-computer interactions, the machine knows the user.
- Suitable for keyword-based IR.
- Phrase matches localized, offering hope for phrase recognition.
- Moore's law favors non-parametric recognizers!

3/16/04

DR&amp;R 2004 Nagy et al.

28

Thank you!

Q?

3/16/04

DR&amp;R 2004 Nagy et al.

29

3/16/04

DR&amp;R 2004 Nagy et al.

30

# EXTRAS

3/16/04 DR&R 2004 Nagy et al. 31

## Our most/least favorite writers

Writer-5 *An emergency brake and is located at either end of the car.*

Writer-6 *An emergency brake and is located at either end of the car.*

Writer-8 *An emergency brake and is located at either end of the car.*

Writer-12 *The risk of you to identify again what the founding fathers actually said.*

Writer-7 *An emergency brake and is located at either end of the car.*

Writer-9 *An emergency brake and is located at either end of the car.*

Writer-10 *I like the small size better than the small room.*

Writer-11 *Have a long time since bill, but there's a lot to do.*

3/16/04 DR&R 2004 Nagy et al. 32

## Comparison with external system (four writers we like) 100-word lexicons

ID \ %	4	5	6	8	Average
Top-1	96.7	95.7	99.7	95.3	96.9
Top-10	98.3	95.7	99.7	97.3	97.9

ID \ %	4	5	6	8	Average
Top-1	94.0	97.3	94.7	94.0	95.0
Top-10	99.0	99.0	99.7	99.0	99.2

3/16/04 DR&R 2004 Nagy et al. 33

## ALLOGRAPHS

Allographs (a) affect a smaller sub-segment of polygrams (ab)

3/16/04 DR&R 2004 Nagy et al. 34

3/16/04 DR&R 2004 Nagy et al. 35

## Lexicon

- First thousand words from the Brown Corpus sorted by frequency of occurrence
- Brown Corpus Characteristics:
  - 43,300 unique words in lower-case letters, apostrophe, and quotation marks
  - 48.68% words with unique set of letters
  - 99.92 words with unique set of bigrams
  - 99.99 words with unique set of trigrams

3/16/04 DR&R 2004 Nagy et al. 36

### Matching Erroneous Feature Graph

3/16/04 DR&R 2004 Nagy et al. 37

### Order Isomorphism and Permutations

- Convert (bipartite) graphs to permutations

- Express constraint in terms of permutations:
 
$$T = (T_1, T_2, \dots, T_n), P = (P_1, P_2, \dots, P_m), n \geq m$$
 Find a longest sequence of pairs:
 
$$\langle T_{i_1}, P_{j_1} \rangle, \langle T_{i_2}, P_{j_2} \rangle, \dots, \langle T_{i_k}, P_{j_k} \rangle$$
 s. t.
  - $(i_1, i_2, \dots, i_k)$  and  $(j_1, j_2, \dots, j_k)$  are both increasing sequences of indices, and
  - $(T_{i_1}, T_{i_2}, \dots, T_{i_k})$  and  $(P_{j_1}, P_{j_2}, \dots, P_{j_k})$  are isomorphic.

3/16/04 DR&R 2004 Nagy et al. 38

### Checking for Permutation Isomorphism

P:	5	11	3	8	16	20	14
V:	0	0	2	1	0	0	2

# elements to the left that are larger.

- V uniquely determines the order of elements in P.

3/16/04 DR&R 2004 Nagy et al. 39

### Host Tree for Matching Permutations

3/16/04 DR&R 2004 Nagy et al. 40

### Search Space Pruning (2)

#### Bounding in a B&B Algorithm

If two sub-permutations are *not* isomorphic, no extension can be isomorphic.

$T = (3\ 2\ 5\ 1\ 7\ 4), P = (1\ 5\ 3\ 4\ 2)$

3/16/04 DR&R 2004 Nagy et al. 41

### B&B (3)

#### Bounding in a B&B Algorithm (cont'd)

$T = (3\ 2\ 5\ 1\ 7\ 4), P = (1\ 5\ 3\ 4\ 2)$

3/16/04 DR&R 2004 Nagy et al. 42

## Premises - 1

- Source-specific recognition is easier
  - In most human-computer interactions, the machine knows the user.
- Many applications require a large, easily extensible vocabulary.
- Lexical context is essential for speech and handwriting recognition.
- Current training methods, based on HMMs are unfriendly and unrepresentative.

3/16/04

DR&amp;R 2004 Nagy et al.

43

## Premises - 2

- Moore's law favors non-parametric recognizers.
- Humans adapt faster than machines but static systems cannot exploit user learning.
- Recognition should improve with use.

3/16/04

DR&amp;R 2004 Nagy et al.

44

## Conclusion

- A new non-parametric approach to whole word recognition
- Improved performance with longer reference string
- May be applicable to:
  - unsegmented phrases
  - speech

3/16/04

DR&amp;R 2004 Nagy et al.

45

## Assumptions

- **Single user**
  - Some training required for each new user of the system
- **Single-word recognition**
  - Perfect segmentation assumed at the word level
- Words from a known **fixed (but extensible) lexicon**
- A sample of handwritten words (**reference-signal set**) and their individual **transcriptions** are available.

3/16/04

DR&amp;R 2004 Nagy et al.

46

## Another Implementation

- Two controls used to eliminate candidate matches:
  1. *Graph size check*: Eliminates lexical graphs that differ substantially in size from the query graph.
  2. *Match size check*: Eliminates candidates with match size substantially different from the best match.
- Matching performed with reference string incremented in steps of 10 words until:
  - a single candidate left, or
  - reference string size of 1000 is reached.

3/16/04

DR&amp;R 2004 Nagy et al.

47

## Types of Errors

1. *False positive (i,j)*: noise is an extra edge at query word string at position  $i$  and reference word at position  $j$ .  
Probability:  $p(e|0)$
2. *False negative (i,j)*: noise is a missing edge at query word string at position  $i$  and reference word at position  $j$ .  
Probability:  $p(e|1)$

3/16/04

DR&amp;R 2004 Nagy et al.

48

## InkLink

Adnan El-Nasan (2003)

For on-line handwriting, we are using the *InkLink* features.

Constrained *localized* polygram matching:  
one unknown word against many reference words,  
using a lexicon of legal words.

Avoids explicit character segmentation  
(of character-based systems)

Avoids training set limitations  
(of word-based systems)

Cannot cope with elastic horizontal distortion

3/16/04

DR&R 2004 Nagy et al.

49

## POLYGRAM MATCHES

Polygram letter and phoneme segments are more distinctive than unigrams, and require no implicit segmentation.

Reference samples of common polygrams (*hundreds*) are easier to obtain than samples of common words (*thousands*).

3/16/04

DR&R 2004 Nagy et al.

50

## InkLink classification algorithm

1. The expected location where the unknown matches the reference words is pre-computed (the number of features in letter of each lexicon and reference word is estimated by least squares).
2. The features matches of the unknown against the reference words are found by string matching.
3. The unknown is hypothesized as each lexicon word in turn.
4. The hypothesis that corresponds best to the expected length and location of the matches is chosen.

3/16/04

DR&R 2004 Nagy et al.

51

## From electronic ink to feature string

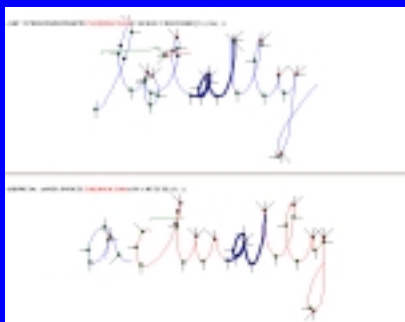


3/16/04

DR&R 2004 Nagy et al.

52

## Feature matching



3/16/04

DR&R 2004 Nagy et al.

53

## Polygram feature match

Unknown query word: "founding"



(0X9j3XaNeEwBxEeNaWwSoXXtWwSoXrTbSutTbNwXxXcXcNaWwSoXjWwSoXTwSoj05%)(NewBwNewSoXcNaWwSoNtj65)

A reference word: "amendment"



(XLXcEeNaWwSoXrTbSutTbNwXxXcXcNaWwSoXjWwSoXTwSoj05%)(NewBwNewSoXcNaWwSoNtj65)



### Matching Match Graphs

Best matching subgraphs preserve Edge Crossings, i.e. find *order-isomorphic subgraphs*

### Results: SIC (simulation only!)

1000 most common words of Brown Corpus

Weighted Noise Model

Probability of Wrong Match:	0.4	0.6	0.8	1.0
% Correct Recognition:	100	100	100	96
Reference String Size:	390	500	600	750
$Mean\left(\frac{Top_1 - Top_2}{Top_1}\right)$	0.20	0.16	0.14	0.13

Adaptation: we can add recognized words to the reference string, as in InkLink

3/16/04 DR&R 2004 Nagy et al. 62

### InkLink --> SIC

SIC replaces the character-length estimation (i.e., implicit segmentation) in InkLink by an *generalized comparison of match graphs*.

It is therefore applicable to patterns that undergo an order-preserving warping.

3/16/04 DR&R 2004 Nagy et al. 63

### SIMULATION ON BROWN CORPUS

Q (% noise)	10	20	30	40
% recognition rate	99.2	99.1	98.3	96.0

3/16/04 DR&R 2004 Nagy et al. 64

### PRINTED TEXT

3/16/04 DR&R 2004 Nagy et al. 65