Semantic Information Extraction with a Thesaurus for Visual Word Recognition

Jonathan J. Hull and An-Tzu Chin
Center of Excellence for Document Analysis and Recognition
Department of Computer Science
State University of New York at Buffalo
Buffalo, New York
hull@cs.buffalo.edu

Abstract

The use of a structural representation for semantic information to improve the performance of a text recognition algorithm is proposed. Semantic constraints are modeled by the graph of connections in a thesaurus, where the thesaurus consists of root words that point to lists of related words. The alternatives produced by a word hypothesization routine are looked up in the thesaurus and activation scores for related words are incremented. In a second pass, the neighborhoods for each word are sorted by the activation scores and a threshold is applied. Since a passage of text is usually about a single topic, the activation scores provide a method of grouping semantically related alternatives. An experimental application of this approach is demonstrated with a word hypothesization algorithm that produces a number of guesses about the identity of each word in a running text. The word recognition alternatives for nouns are used in two recursive thesaurus lookups. The resulting activation values are used to threshold the number of alternatives that can match each word. It is shown that a reduction of 12 to 20 percent in average neighborhood size can be achieved with a one percent error rate.

1. Introduction

Text recognition algorithms often process only images of isolated characters. This is sometimes followed by a post-processing step that uses information from a dictionary of allowable words to correct recognition errors. This approach can provide high performance for good quality images. For example, 99.5 percent correct character recognition rates are commonly reported for commercial character recognition devices on clean images. However, even this level of performance still implies that a typical page of text containing about 4000 characters or approximately 800 words would still contain about 20 errors. More importantly, this performance would be much worse if degraded images were input such as facsimile documents or multiple generation photocopies.

appeared in Advances in Structural and Syntactic Pattern Recognition, H. Bunke (editor), World Scientific, Singapore, 1992, pp. 342-351.

A computational model for word recognition has been proposed that overcomes some of the constraints of other methodologies [5]. The model suggests that to achieve high levels of performance in text recognition for a range of input qualities it may be necessary to understand the text while recognizing it. One part of the understanding process is an analysis of the meaning of the text. The underlying meaning of text has been utilized in an innovative way to improve performance in a restricted domain by allowing only alternatives that were sensible in context [1]. The use of definitional overlap in machine-readable dictionaries and word collocations as semantic representations to improve the performance of a handwriting recognition system has also been discussed [7, 2]

This paper proposes to model the semantic relationships between words by the network of connections in a thesaurus. The thesaurus is represented by a directed graph in which root nodes point to related words. These words can also be root words that in turn point to other words, and so on. This structural representation is applied to text recognition by first using a word recognition algorithm to supply a number of alternatives for the identity of each word. The alternatives for the words in a passage of text are then recursively looked up in the thesaurus. Counters associated with each thesaurus word are incremented when the words are encountered in the lookup process. Semantically related alternatives in the neighborhoods from a passage of text are thus reinforced and unrelated words are suppressed. A threshold is applied to the scores to remove unrelated words from neighborhoods.

The rest of this paper outlines the framework of the text recognition algorithm. This is followed by a discussion of the thesaurus structure and its usefulness. A method of simulating word recognition performance is presented and experimentation with different levels of thesaurus lookup is discussed.

2. Text Recognition Algorithm

The recognition algorithm that incorporates the thesaurus constraint contains the three steps shown in Figure 1. The input is a sequence of word images $w_i, i = 1,2, \cdots$. The hypothesis generation stage computes a group of possible identifications for w_i (called N_i or its *neighborhood*) by matching a feature representation of the image to the entries in a dictionary. The global contextual analysis phase uses information about other words that have been recognized, such as their semantic classification, to constrain the words that can be in N_i. The output is a neighborhood N_i^* of reduced size. The hypothesis testing phase uses the contents of N_i^* to determine a specific set of feature tests that could be executed to recognize w_i. The output of hypothesis testing is either a unique recognition of w_i or a set of hypotheses that contain the word in the image.

appeared in Advances in Structural and Syntactic Pattern Recognition, H. Bunke (editor), World Scientific, Singapore, 1992, pp. 342-351.

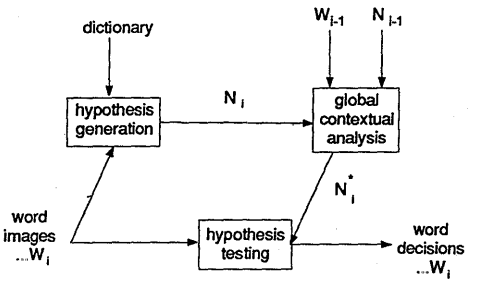

Figure 1. The text recognition algorithm.

3. Semantic Model: The Thesaurus

The thesaurus is a directed graph of words. An example of four entries in the thesaurus, for the words AARDVARK, ABACK, ABAFT, and ABANDON, is shown in Figure 2.[1] The thesaurus contains entries for 25,072 words. Each entry provides a root word and a list of related words. A minimum of two, a maximum of 257, and an average of 49.2 related words are in each entry. There are 50,357 different non-root words in the thesaurus.

The potential usefulness of the thesaurus in text recognition is illustrated by looking up all the words in a sample of text, accumulating the counts, and observing their distribution of values. The result of applying this process with two levels of lookup to 2000 words of a story about golfing[2] is illustrated in Table 1. The top 40 scores and the associated words are shown. It should be noted that no scores were accumulated for the true words. Thus any natural bias is eliminated and the pure effect of the thesaurus is given.

It is very interesting to observe that key terms about golf are very frequent. For example, *golf* COURSE, ROUND *of golf*, *make the* CUT, *driving* RANGE, HOLE *in one*, are all near the top of the ranking. Numerous other similarities are also observed.

[1] The commercially available Moby Thesaurus was used in this project.

[2] Alfred Wright, "A duel golfers will never forget," Sports Illustrated, 64, pp. 18-21, April 17, 1961.

appeared in Advances in Structural and Syntactic Pattern Recognition, H. Bunke (editor), World Scientific, Singapore, 1992, pp. 342-351.

AARDVARK
ARTIODACTYLA, PRIMATES, MAMMAL,

ABACK
SHORT, SUDDENLY, UNAWARE, UNAWARES, UNEXPECTEDLY, UNPREPARED,

ABAFT
AFT, AFTER, ASTERN, BACK, BEHIND,

ABANDON
ABANDONMENT, ABJECTION, ABJURE, ABORT, ABSCOND, AGITATION, ARDENCY, ARDOR, BREAK, CANCEL, CAPITULATE, CARE, CARELESSNESS, CAST, CEASE, CEDE, CHUCK, CORRUPTEDNESS, CORRUPTION, CORRUPTNESS, CRAZE, DEBASEMENT, DECADENCE, DECADENCY, DEEP-SIX, DEGENERACY, DEGENERATENESS, DEGENERATION, DEGRADATION, DELIRIUM, DEMORALIZATION, DEPRAVATION, DEPRAVEDNESS, DEPRAVITY, DESERT, DESIST, DISAPPEAR, DISCARD, DISCONTINUE, DISJOIN, DISMISS, DISOBEDIENCE, DISQUIET, DISQUIETUDE, DISREGARD, DISREGARDFULNESS, DISSOLUTENESS, DISTRESS, DISTURBANCE, DISUSE, DROP, DUMP, DWINDLE, EAGERNESS, EASE, EBB, ECSTASY, EIGHTY-SIX, ELIMINATE, ENCHANTMENT, END, EVACUATE, EXCESS, EXCESSIVENESS, EXCITEMENT, EXTRAVAGANCE, EXTRAVAGANCY, EXUBERANCE, FAIL, FERVENCY, FERVIDNESS, FERVOR, FIRE, FORGET, FORGETFULNESS, FORGIVE, FORGIVENESS, FORGO, FORSAKE, FORSWEAR, FREEDOM, FRENZY, FUN, FUROR, FURY, HALT, HEARTINESS, HEAT, HEATEDNESS, HEEDLESSNESS, HOLD, HYSTERIA, IMMODERACY, IMMODERATENESS, IMMODERATION, IMPASSIONEDNESS, IMPETUOUSNESS, IMPULSIVENESS, INCONSIDERATENESS, INCONSIDERATION, INCONTINENCE, INDISCIPLINE, INORDINATENESS, INQUIETUDE, INSUBORDINATION, INTEMPERANCE, INTEMPERATENESS, INTENSITY, INTENTNESS, INTOXICATION, IRREPRESSIBILITY, JETTISON, JILT, JUMP, JUNK, LAXITY, LAXNESS, LEAVE, LIBERTY, LICENSE, LICENTIOUSNESS, LOOSENESS, MADNESS, MALINGER, MAROON, MISS, MUTINY, NEGLIGENCE, NERVOUSNESS, NIMIETY, NONCOERCION, OBLIVION, OMIT, ORGASM, ORGY, OVERINDULGENCE, PASSION, PASSIONATENESS, PERMISSIVENESS, PERORATE, PERTURBATION, PLAY, PRETERMIT, QUIT, QUITCLAIM, RAGE, RAPTURE, RAVISHMENT, REBUFF, REFRAIN, REJECT, RELEASE, RELINQUISH, REMOVE, RENOUNCE, REPEL, REPUDIATE, REPULSE, RESIGN, RESOLUTION, RETIRE, RETRACT, RETREAT, RIOTOUSNESS, SACRIFICE, SCRAP, SHIRK, SKIP, SLACK, SLOUGH, SPARE, SPIRIT, SPORT, STAY, STOP, SUBMIT, SUBSIDE, SURRENDER, TACTLESSNESS, TERMINATE, THOUGHTLESSNESS, TRANSPORT, TRIFLE, TROUBLE, TURPITUDE, UNEASINESS, UNPREPAREDNESS, UNREADINESS, UNRESERVE, UNRESTRAINT, UNRULINESS, UNTHINKINGNESS, UPSET, VACATE, VANISH, VEHEMENCE, VEXATION, WAIVE, WANE, WARMTH, WILDNESS, WITHDRAW, YIELD, ZEAL,

Figure 2. Four thesaurus entries

appeared in Advances in Structural and Syntactic Pattern Recognition, H. Bunke (editor), World Scientific, Singapore, 1992, pp. 342-351.

COURSE	3423	ROUND	3187	CUT	2445	LINE	2427
RANGE	2360	HOLE	2330	POINT	2121	SET	1990
TURN	1985	CRACK	1885	PLACE	1862	CHECK	1794
SPACE	1790	MEASURE	1708	MARK	1687	PERIOD	1625
SCALE	1608	STEP	1607	ORDER	1571	FIELD	1561
STAGE	1548	RANK	1526	BEAT	1493	LEVEL	1393
OPENING	1383	NOTCH	1374	CYCLE	1350	PIT	1348
SHADOW	1334	RACE	1318	BANK	1304	GROUND	1304
DESCENT	1292	STRING	1290	PASSAGE	1274	CIRCUIT	1260
INTERVAL	1239	GALLERY	1237	COVER	1227	WAY	1224

Table 1. Thesaurus activation values after two levels of lookup

4. Algorithm

A statement of the algorithm for applying the thesaurus constraint to filtering neighborhoods output by the word recognition process is below:

1. Initialize counters for every thesaurus word (50,357)
2. For every word in every neighborhood:
   Recursively increment activation counters for lookup(word).
3. Rank the neighborhoods of content words (nouns) by:
   Confidence(word) = counter(word) / sum of counters in neighborhood
4. Reduce neighborhoods by thresholding the confidence value.

The function lookup(word) returns the list of related words from the thesaurus.

5. Example

An example of applying the thesaurus constraint is shown in Figure 3. An original sentence and the neighborhoods for the nouns in the sentence are given. The activation counters and confidence values are shown for each word after two levels of lookup in the thesaurus. The neighborhoods for each word in the story about golfing mentioned earlier were calculated by a simulation of the word recognition process. In this case, a threshold of 0.25 would yield correct answers for the first and third words. All other entries in those neighborhoods would be eliminated. However, the second neighborhood would contain two words, neither of which were correct.

appeared in Advances in Structural and Syntactic Pattern Recognition, H. Bunke (editor), World Scientific, Singapore, 1992, pp. 342-351.

original sentence

```
The final round was played on the golf course.
```

neighborhoods ranked by confidence values

```
round = round (16077, 0.49), tunnel (5056, 0.15), mood (2909, 0.09), counsel (2232, 0.07),
        removal (2153, 0.07), enamel (2080, 0.06), general (2067, 0.06), mantel (221, 0.01),
        graveyard (85, 0.00).

golf = gulf (6740, 0.69), gully (2617, 0.27), golf (368, 0.04).

course = course (19523, 0.26), range (16015, 0.22), game (7005, 0.10), master (6948, 0.09),
         tenor (5677, 0.08), center (5581, 0.08), genus (4527, 0.06), name (4120,0.06),
         terrace (3991, 0.05).
```

Figure 3. Application of thesaurus constraints to an example sentence.

6. Experimental Investigation

Experimental tests were conducted to determine the ability of the thesaurus constraint to reduce the word candidates that match any image. Hypotheses were generated for the words in a test sample of text and the thesaurus ranking algorithm was applied. Two levels of performance were calculated: the maximum amount the neighborhoods could be reduced without incurring any errors, based on the ranking of words provided by the thesaurus, Also, the reduction in neighborhood size possible with a fixed threshold on the confidence value was determined.

Performance was measured by calculating the average neighborhood size per text word before and after the application of the thesaurus constraint. This statistic is defined as:

$$ANS_t = \frac{1}{N_w} \sum_{i=1}^{N_w} ns_i$$

where N_w is the number of words in the test sample and ns_i is the number of words in the neighborhood for the i^{th} word in the text. The *error rate* is the percentage of words with neighborhoods that do not contain the correct choice after the application of syntax.

appeared in Advances in Structural and Syntactic Pattern Recognition, H. Bunke (editor), World Scientific, Singapore, 1992, pp. 342-351.

6.1. Text Database

A soft copy (ASCII) text sample known as the Brown Corpus was used for the experiments [6]. This text was chosen because it is large (over 1,000,000 words of running text) and every word is tagged with its syntactic class. The corpus is divided into 15 subject categories or genres that span a range from newspaper reportage to Belles Lettres. There are 500 individual samples of running text in the corpus and each one contains approximately 2000 words. The number of samples in each genre differs depending on the amount published in that area at the time the corpus was compiled.

6.2. Hypothesis Generation Algorithm

The operation of the hypothesis generation algorithm was simulated by calculating the feature description for a word from pre-defined features for the letters in the word. All the words in a dictionary with the same feature description were used as its neighborhood. Two feature descriptions that produce different neighborhoods were used to demonstrate the effect of neighborhood size on performance.

The feature descriptions are specialized for lower case characters because the experimentation is restricted to text written in lower case. The first feature description includes vertical bars of different heights, dots, and empty spaces. These features were chosen because they can be reliably computed from images of text even if the characters touch one another [4]. When a feature description is applied to a word it yields a symbolic representation that would correspond to the sequence of occurrence of the features in an image of the word. Thus, both "me" and "may" have a symbolic representation 22221 and the neighborhood of "me" is {"me", "may"}.

The second feature set includes all the features in the first description plus the holes produced by topological containment. The addition of this feature provides a finer discrimination in neighborhood calculation, i.e., smaller neighborhoods.

6.3. Experimental Results

The thesaurus algorithm was applied to correct the simulated text recognition results produced by the two models for hypothesis generation described earlier. The 259 nouns in subset A38 of the Brown corpus were selected as test data. Two neighborhoods were calculated for each word and two levels of thesaurus lookup were performed. The lexicon for the entire corpus was used. This provided an average neighborhood size of 1.57 for the first feature description and 17.50 for the second feature description. A neighborhood size of approximately 15 words has been shown to be sufficient to achieve better than a 98 percent recognition accuracy within the neighborhood [3]. The best possible reduction in neighborhood size (*oracle* performance) was determined. This provides a measure of the upper bound in neighborhood reduction that could be achieved without incurring any errors. The results are shown in Table 2.

appeared in Advances in Structural and Syntactic Pattern Recognition, H. Bunke (editor), World Scientific, Singapore, 1992, pp. 342-351.

no. lookups	feature set 1	feature set 2
1	13.48%	35.47%
2	18.71%	42.60%

Table 2. Oracle performance: neighborhood reduction with 0% errors

The usefulness of the thesaurus in a running system that employed a global threshold on the confidence values over all the neighborhoods was also demonstrated. The same experimental procedure was applied to sample A38 and the reduction rates and error rates incurred at different thresholds were determined.

The results of this experiment are shown in Table 3. It is seen that using the first feature description and one level of lookup in the thesaurus, a neighborhood size reduction of about ten percent could be achieved with an error rate of less than half a percent. This performance is relatively small because the input neighborhoods on average contained only 1.57 words. The performance with the second feature description was more indicative of the value of the thesaurus information. A reduction of 14.82 percent was achieved with an error rate of 1.54 percent. When two levels of lookup were used, this was improved to a 19.59 percent reduction with a 0.77 percent error rate.

appeared in Advances in Structural and Syntactic Pattern Recognition, H. Bunke (editor), World Scientific, Singapore, 1992, pp. 342-351.

thresh	feature desc. 1				feature desc. 2			
	1 lookup		2 lookups		1 lookup		2 lookups	
	reduce	error	reduce	error	reduce	errors	reduce	errors
0.01	0.00	0.00	0.60	0.00	14.82	1.54	19.59	0.77
0.02	2.62	0.00	0.80	0.00	21.22	3.09	29.03	3.09
0.03	4.02	0.00	0.80	0.00	28.10	13.51	35.21	7.72
0.04	4.02	0.00	1.21	0.00	32.27	15.44	38.76	15.06
0.05	4.02	0.00	2.01	0.00	36.00	16.60	41.96	17.37
0.06	4.23	0.00	2.01	0.00	37.52	17.37	44.47	19.31
0.12	5.84	0.36	7.65	0.00	44.80	26.25	52.55	29.34
0.14	7.65	0.36	9.26	0.36	46.15	27.41	54.29	35.52
0.15	8.45	0.36	9.26	0.36	47.05	29.34	54.75	37.84

Table 3. Percent reduction in *ANS*, and error rate at various thresholds on the look-up score.

7. Discussion and Conclusions

An algorithm for incorporating semantic constraints in word recognition that uses a graph structured thesaurus to group word alternatives was presented. A recursive lookup in the thesaurus was shown to yield a maximum potential reduction in neighborhood size of 43 percent when two levels of lookup were performed. Implementation of a simple threshold on the same data provided only a 20 percent reduction with less than a one percent error rate.

Future work will explore the usefulness of deeper thesaurus lookups. A more intelligent method of applying a threshold will be considered to get closer to the 43 percent reduction level. Incorporation of the semantic weights provided by the thesaurus lookup in a statistically based syntactic analysis will also be considered.

appeared in Advances in Structural and Syntactic Pattern Recognition, H. Bunke (editor), World Scientific, Singapore, 1992, pp. 342-351.

References

1. H. S. Baird and K. Thompson, "Reading Chess," *IEEE Transactions on Pattern Analysis and Machine Intelligence 12* (1990), 552-559.
2. L. J. Evett, C. J. Wells, F. G. Keenan, T. Rose and R. J. Whitrow, "Using linguistic information to aid handwriting recognition," *International Workshop on Frontiers in Handwriting Recognition*, Bonas, France, September 23-27, 1991, 303-311.
3. T. K. Ho, J. J. Hull and S. N. Srihari, "Combination of Structural Classifiers," *IAPR Workshop on Syntactic and Structural Pattern Recognition*, Murray Hill, New Jersey, June 13-15, 1990, 123-136.
4. J. J. Hull, "Hypothesis generation in a computational model for visual word recognition," *IEEE Expert 1*, 3 (Fall, 1986), 63-70.
5. J. J. Hull and S. N. Srihari, "A computational approach to visual word recognition: hypothesis generation and testing," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami Beach, Florida, June 22-26, 1986, 156-161.
6. H. Kucera and W. N. Francis, *Computational analysis of present-day American English*, Brown University Press, Providence, Rhode Island, 1967.
7. T. G. Rose, L. J. Evett and R. J. Whitrow, "The use of semantic information as an aid to handwriting recognition," *First International Conference on Document Analysis and Recognition*, Saint-Malo, France, September 30 - October 2, 1991, 629-637.

appeared in Advances in Structural and Syntactic Pattern Recognition, H. Bunke (editor), World Scientific, Singapore, 1992, pp. 342-351.

Series in Machine Perception and Artificial Intelligence – Vol. 5

ADVANCES IN STRUCTURAL AND SYNTACTIC PATTERN RECOGNITION

Proceedings of the International Workshop on Structural and Syntactic Pattern Recognition

Bern, Switzerland August 26 – 28, 1992

Edited by

H. Bunke
Institut für Informatik und Angewandte Mathematik
Universität Bern
Switzerland

World Scientific
Singapore • New Jersey • London • Hong Kong

appeared in Advances in Structural and Syntactic Pattern Recognition, H. Bunke (editor), World Scientific, Singapore, 1992, pp. 342-351.