

Application of Pictographic Recognition Technology for Spotting Handwritten Chinese Words

Donald T Gantz, PhD **John J Miller, PhD**

George Mason University
4400 University Drive
Fairfax Virginia 22030
(703) 993-1511

dgantz@gmu.edu

jmiller@gmu.edu

Mark A Walch

The Gannon Technologies Group
1000 North Payne Street
Alexandria, Virginia 22314
(703) 373-1962

mwalch@gannontech.com

Abstract

Pictographic Recognition Technology is a Graph-Theory-based methodology for the detection and extraction of specified words or groups of words from both handwritten and machine printed document collections. This technique converts written and printed forms into mathematical graphs and draws upon key properties of these graphs such as topology and geometric features to locate graphs that respond to specified search terms. In its initial implementation, Pictographic Search Technology has functioned by identifying individual characters in strings of handwritten script—both English and Arabic.

Handwritten Chinese presents the opportunity to exploit a key capability of Pictographic Recognition: detection of embedded forms. These embedded forms reflect graphical structures that are consistent for particular Chinese words—similar to the concept of “radicals”. This paper focuses on current research toward using Pictographic Recognition techniques to spot handwritten Chinese words.

1 Introduction

Pictographic Recognition offers a means for spotting handwritten Chinese words within large document collections. The mathematical foundation for Pictographic Recognition is Graph Theory and the technique entails the use of graph-based pattern matching techniques to detect the “graphical signature” of objects contained in images of documents. The objects can be single characters, words, groups of words or characters, signatures and marks, symbols and virtually any other graphical form. Every written object contains a graphical signature constructed from graphs. These graphs may either constitute the full written object or be embedded within it. Pictographic Recognition permits these graphs, both full and embedded, to be automatically isolated, classified and linked to known

graphs representing alphabetic characters, groups and parts of character components, parts of signatures and other forms. And, in the case of Chinese, these graphs can represent major components within the written word that form the basis for word spotting and recognition. To date, Pictographic Recognition has been implemented as a Prototype Application that performs word searches based on pictographic signatures. Current languages where Pictographic Recognition has been implemented include machine printed, hand printed and cursive English as well as machine printed and handwritten Arabic.

Chinese words typically consist of very complex graphical forms containing two or three times the number of edges and vertices that can be found in other languages such as English or Arabic. However, their strong pictographic roots suggest Pictographic Recognition should prove to be a useful tool for word identification.

Research to-date indicates that the complex graphical structure offers an advantage in identifying machine-printed Chinese words where the printed form offers such a level of consistency that graph complexity can be used as an identifier. That is, some printed Chinese forms can be recognized or spotted purely through the topology of their “whole word” graphs. However, since handwritten forms exhibit a much higher degree of variability than machine-printed forms, the complexity of Chinese poses an extremely daunting challenge both to computer-based recognition and even to automated techniques for spotting words of interest.

The authors believe spotting words of interest in Chinese writings often follows a model akin to “signature identification”. That is, a person’s signature is complex and almost never written exactly the same way twice. However, there are similar elements within each written signature that enable the

signature to be identified. Similarly, the Chinese word written by the same individual will be different—sometimes significantly different—with each new writing. And, the same word written by multiple individuals will exhibit even more differences. However, as is also true of signatures, different writings of the same Chinese word will retain a kernel of key graphical information that permits recognition. This graphical information is what Pictographic Recognition can detect. It must be emphasized that no single kernel can be expected to be consistent for all forms of the same word. However, it is believed a manageable set of kernels can be identified that can provide broad coverage across variant word forms sufficient to support word spotting—and ultimately support recognition. The current research focuses on isolating and identifying kernels of graphical information embedded in Chinese words. The objective is to build sets of these embedded forms that can be used for recognition-based word spotting.

Because of their pictographic roots, graph based forms are the basis upon which the Chinese language is predicated. For purposes of Pictographic Recognition, these basic building blocks of Chinese writing, the “radicals”, are simply embedded graphs. It is believed that sets of embedded graphs can be extracted from Chinese writing and that the topology and features of these graphs will contribute to identification of the word in which they were embedded. These embedded graphs may include actual radicals plus other graphical forms that behave like radicals. The objective of the present research is to find forms that can be automatically extracted from multiple writings from different authors. Once isolated, these forms must be able to identify reliably the words from which they were extracted while minimizing confusion with other words.

Pictographic Recognition shares common roots with Optical Character Recognition (“OCR”), but it is a fundamentally different approach. Pictographic Recognition looks for the “Pictographic Signature” of words in their native form. In this context, Pictographic Recognition is somewhat similar to Optical Word Recognition (“OWR”) but differs significantly from OWR in its ability to evaluate the actual characters and character groupings that comprise the unknown words. The greatest distinction of Pictographic Recognition from OCR and OWR processes is the ability to look both at the full form of a character or word as well as the ability to “drill down” into embedded graphical forms.

The discussion that follows will focus on the ability of Pictographic Recognition to identify Chinese words through identification of embedded graphical

forms. At the heart of this recognition process is the ability to encode both the features and topology of a Chinese word into a form that can be used for rapid identification. The technique herein described is strongly rooted in an algorithmic approach, based on Graph Theory, which treats handwriting and printed text as mathematical graphs. The distinction that applies to Chinese writing is that the principles successfully applied to other languages with linear character relationships will now be extended to “two-dimensional” pictographic structures.

2 Conceptual Framework

Graph Theory is a branch of Mathematics that focuses on representing relationships as line diagrams containing nodal points and the linkages among these points. In graph terminology, the nodes are referenced as “vertices” and the links as “edges”. Graphs are an effective way to represent written language since graphs can be created directly from the pen strokes used to compose letters and words. Words, characters and numerals take their form as graphs written on paper assuming distinctive shapes representing the letters of the alphabet as well as numerals and punctuation. The edges and vertices of these written graphs connect and cross and are straight or curved. Graphs accurately capture the essence of written language since they contain both the topological structure and geometric information sufficient to replicate complete written forms. Graphs are the foundation of written language and, therefore, a highly efficient and accurate tool within which to conduct searches, irrespective of languages as well as handwritten or machine generated formats.

Graphs contain all the information extracted from writing condensed into a concise mathematical format that is highly computable. Within graphs, this information takes two forms. The first form is the graph’s topology that can be seen in the connectivity among the major graph components. The way pen strokes are crossed and connected is the framework for graph topology. Topology is the structure of the graph. The second form of information contained in graphs is the geometry of the graph. Geometry is expressed in terms of distances, angles and characteristics of graph components. The depth or shallowness of a curve, the distance between line crossings or the sharpness of angles are all examples of graph geometry. Geometry characterizes the shape of the graph. Collectively, topology and geometry account for the structure and shape of graphs. The topology and geometry of graphs are also quite computable. That is, they can be expressed as data to support computer-based processes that can be performed on graphs such as indexing and searching. The foundation of Pictographic Recognition is an algorithm that describes graph topology as a numeric

code. Any two graphs with the same structure will generate the same code. Any two graphs generating the same code are said to be isomorphic.

Although the details of how this code is constructed are beyond the scope of this paper, the method for generating the code involves reordering a graph's adjacency matrix based on connectivity of the vertices within the graph. This method will create identical adjacency matrices from graphs that are isomorphic. The vertex with the highest level of connectivity will always be placed first in the new graph ordering. This vertex is referenced as the "Prime Vertex". The Prime Vertex is important because it provides the basis for vertex ordering during key construction. It also provides a readily identifiable reference point within the graph that can be used as the origin for distance and angular measurements. The reference point afforded by the Prime Vertex provides a foundation for encoding the shape of a graph which is discussed later in this paper.

Linked to the graph topology and its attendant numeric code is graph geometry. Graph geometry can also be expressed as a feature vector used to compare graphs. A feature vector is a multi-dimensional expression containing the multitude of measurements that can be extracted from graphs. The topology of graphs can be coupled with feature vectors as a combined data structure. The result is a structure that is very computable while offering an effective way to use graphs as surrogates for characters and words contained in the images of documents.

The computational engine for Pictographic Recognition is an automated method for identifying and matching isomorphic graphs and storing these graphs in a database. Graphs are isomorphic when they are structurally identical—with the same number of edges and vertices connected in the same way—although they may appear to be different. Two graphs are considered isomorphic when there exists a one-to-one correspondence between their internal structures. That is, they have the same number of edges and vertices connected to form exactly the same topology.

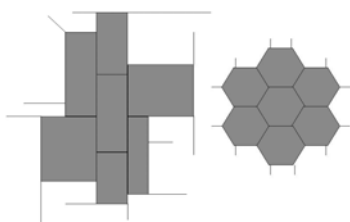


Figure 1: Illustration of two Isomorphic Graphs with different features.

Figure 1 illustrates this point. Although the graphs appear to be quite different, they are structurally identical: isomorphic. They appear different because their geometry is different. The topology is the same, but the angles and distances are different.

Figure 2 illustrates the concept of Isomorphic Graphs as it applies to characters from handwritten English. In this figure, each row shows three instances of the letter "a" written as the same isomorphic graph class. The numbers on the left hand side of the figure are the code names derived by the Pictographic Recognition process for each class of graph. These codes will always be the same for graphs having the same topology—isomorphic graphs. The class labeled "2;192" consists of graphs built from a loop and one edge. Class "4;112.0" contains graphs with three edges only connected at a central vertex. These typically characterize the letter "u" and the letter "a" written to be open at the top. Class "4;98.0.64" encompasses characters with a leading edge, an enclosed area—in graph terminology this is called a "face" -- and a trailing edge.

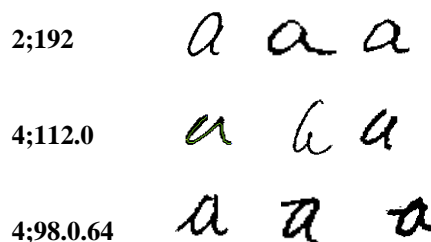


Figure 2: Sample letters "a" for three different graph isomorphic classes.

Figure 3 provides another English language example showing how graph isomorphic classes transcend individual letters. This figure presents letters "a" and "e" for isomorphic graphs classes coded "4;112.0" and 4;98.0.64".

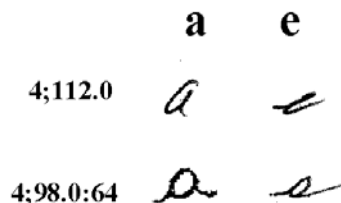


Figure 3: Example of letters "a" and "e" sharing the same isomorphic graph.

In English and Arabic implementations, Pictographic Recognition focuses on locating and identifying individual characters within handwritten words (machine printed matter can be considered to be

“perfect handwriting”). In these cases embedded character forms are encountered in a linear fashion. The discussion in this paper looks at Pictographic Recognition applied in a different manner—as a means for identifying embedded forms that span 2-dimensional shapes where there is both consistency and variability among embedded forms.

The premises underlying mining embedded forms can be summarized through the following points.

- Every line figure can be represented by a graph.
- Every graph can be identified by a unique code that represents its topology.
- Graphs fully capture the topology of written forms as well as all their embedding.
- The full topologies and all embeddings can be represented by codes related to graph isomorphism.
- Geometric feature information can also be stored in graphs.
- These geometric features can also be encoded and “bundled” with the topology information.
- Topology coupled with geometric features creates a very powerful means for classifying graph shapes.
- These classified graphs can be used to identify the written forms from which they were extracted.

3 Embedded Forms

The key to Chinese word spotting is the ability to isolate and to recognize graphs embedded within the graphs that constitute the actual word. The concept of embedded graphs contained in written forms is shown in Figure 4 which illustrates the letters “I”, “L” and “F” embedded in the letter “E”.



Figure 4: Examples of characters embedded in the letter “E”.

Given a particular graph and a set of graphs of interest, Pictographic Recognition can isolate all embeddings that match the graphs of interest. Figure 5 shows a handwritten Chinese character and three potential embedded forms contained within the

character. These are only a subset of the numerous embeddings that can be detected within these words.



Figure 5: Chinese character and three sample embedded forms.

Three factors relate to embedded form detection:

1. Graph Isomorphism
2. Shape Description
3. Embedding Percentage

The Graph Isomorphism represents the specific topology of the embedded graph as described by the unique code identifier provided through Pictographic Recognition. The Shape Description is an encoded means for articulating the geometry of the graph. This paper presents an angle-based encoding method that articulates the form of the graph as a series of directions from the Prime Vertex. And, the Embedding Percentage shows what proportion of the overall word graph the embedded graph occupies. Embedding percentage is relevant in two distinct cases: modeling and testing. In modeling, the embedding percentage determines what exemplars are best suited for reference purposes. In testing, the embedding percentage determines what proportion of an object must match a reference exemplar to determine a match does indeed exist.

These three parameters provide a means of defining and exploiting embedded graphs for recognizing handwritten Chinese.

4 A Chinese Primer

Unlike most of the world's scripts, which evolved towards syllabaries or alphabets, the Chinese script draws on its pictographic and ideographic origins to form picto-phonetic means of representing words and components of words.

Most of the world's scripts fall along a continuum, with purely pictographic means of representing a language on one end and purely phonetic means on the other. Many of the world's older scripts, such as Cuneiform, Egyptian hieroglyphics, and Chinese words, are picto-phonetic. This means that their characters represent both sound and meaning to different degrees.

Chinese is unique among the world's major scripts in use today in that it has several purely pictographic characters. This means that the character is a picture of the word it represents. For example, the character for the Chinese word *ren* “person”, 人 depicts the legs and body of a person. The character for the word *shan* “mountain”, 山 depicts the outlines of a mountain. It is important to remember that pictographic characters are not always immediately apparent. For example, the Chinese word *shui* “water”, 水 is a picture of three streams flowing together.

Anyone who has played charades or “Pictionary” knows that a purely pictographic means of representing words does not go very far, since certain concepts are too abstract. Pictographically, concepts such as “good,” “big,” and “middle” can only be represented using pictures which in some way demonstrate the concept. For example, the word for “big”, 大 *da* is a picture of a man 人 holding his arms out, describing something that is big. Interestingly, combining the words for “wife”, 女 *nü*, and “children”, 子 *zi* creates a combined pictograph for “good”, 好 *hao*. Other characters are more abstract. The character for “middle”, 中 *zhong*, is a line drawn through an indeterminate object which in machine printed form looks like a square or rectangle.

These methods can produce several hundred characters. However, they are not sufficient to represent the vocabulary of an entire language. The example of charades is helpful here – in the same way that someone would act out a word that the target word sounds like, and then act out this word's meaning, most Chinese words combine two characters, one representing its meaning and the other giving a clue as to how to pronounce it.

The character representing its meaning, or the radical, will be one of 214 characters which represent broad semantic categories. For example, the radical from the word for “water”, 水 *shui* will have to do with liquid, water, and fluidity. The character for “eye”, 目 *mu* will have to do with sight, or vision. As a radical, *shui* becomes three dots on the left of the character. Using this radical, we can derive the characters 洲 *zhou* “island”, 沉 *kuang* “cold water”, 油 *you* “oil”, and 游 *you* “swim”.

It also must be remembered that the radical and phonetic components do not necessarily occur on the left and right sides of the character, respectively. In some cases, the alignment can be vertical, as in 宙 *zhou* “universe”, with the radical for a roof.

5 Radicals and Pseudo-Radicals

Since radicals are the basic building blocks of written Chinese, they are graphical forms that can be expected to hold some level of consistency across different Chinese words. Within the contexts of characters-as-graphs, Radicals can be viewed as embedded forms within the graph of a complete Chinese word. Since Pictographic Recognition is capable of detecting graphs embedded in other graphs, it is very consistent with the concept of Radicals in written Chinese.

The present study focuses on the concept of embedded graphs as “radicals” as a substitute for actual semantic values in written Chinese. That is, it is possible to identify automatically embedded graph forms that transcend numerous instances of the same written Chinese character. These forms are embedded graphs that perform the function of a radical and may or may not be the true radical from the Chinese word. The premise is that Pictographic Recognition can detect certain embedded forms inside Chinese words that a computer can use for recognition in a manner similar to the way a native Chinese speaker would “decode” the radicals and other stroke information into the actual word meaning.

These “pseudo-radicals” are empirically extracted graph forms that can be found in numerous instances of the same Chinese character. It must be stressed that handwritten Chinese, like other handwritten forms, exhibits a great deal of variability in topology and geometry so that it should not be expected that a single pseudo-radical (or any embedded graphical form) will encompass the majority of instances of any particular character. However, with enough training samples it is postulated that a set of common consistent graph forms can be identified.

6 Methodology

The methodology for evaluating the concept of pseudo-radicals involves isolating embedded graph forms and classifying these forms. A simple classification method can be used involving the previously discussed “Prime Vertex” from the isomorphic graph key generation process and Feature Angles (FAs) from this vertex to (1) all other vertices in the graph and (2) to certain points on each edge within the graph. The Feature Angles were expressed as three digit degree values from “000” to “359” and concatenated into a numeric string. The order of the directions was based on the ordering of the vertices used to build the isomorphic key in the Pictographic Recognition process. The string of Feature Angles serves as a very simple compact “shape code” in the

spirit of “chain codes” that are commonly used for shape definition.

Figure 6 shows the Chinese word “big” (大 *da*) structured as a graph with the Prime Vertex in the center with Feature Angle pointers emanating as lines toward other vertices. This basic form can be translated into a shape code based on the composite directions.

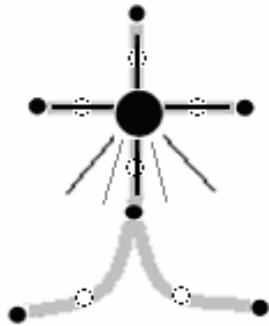


Figure 6: Chinese word “big” showing Prime Vertex and directions to other vertices (black dots) and interior points (white dots).

For purposes of illustration and omitting detailed discussion of the actual methodology for producing the order of the vertices, the directions from the Prime Vertex in Figure 6 would be the following (Our system uses 000 to represent due east, 090 to represent due south, and so forth.): 090 (due south), 000 (due east), 270 (due north), 180 (due west), 115 (southwest), and 065 (southeast) for the vertices and 090, 000, 270, 180, 100, 080 for the interior points, producing the set of angles: 090, 000, 270, 180, 115, 065, 090, 000, 270, 180, 100, 080.



Figure 7: Examples of two different characters with same graph isomorphism distinguished by Feature Angles.

To help the reader understand the value of the interior point angle, consider the two characters in Figure 7 above. The character on the left is a *ge* while the character on the right is a *ren*. The angles for the *ge* are 048, 078, 083, 115, 081 beginning with the vertex at the top of the character. The angles for the *ren* are 041, 162, 255, 045, 208 beginning with the vertex at the center of the character. (The reader may wish to verify how the angles are calculated.) Both characters are represented by the same isomorphism,

namely 4:64.128, which represents “two lines”. However, the angles to the interior points clearly distinguish between the *ge* with the bent line and the *ren* with straight lines. (In fact, the interior point for a bent line is placed at the point of maximum bending.) Using the angles to the interior points proves to be a powerful tool in the recognition of these characters.

7 Analysis of Chinese Sample Writings

Given the inherent complexity in handwritten Chinese, the Authors focused initially on the more simple character forms under the premise that identification of the complex forms will entail identifying a preponderance of embedded simple forms. To obtain a mixture of simple and complex forms, the “20 most common Chinese words” were selected for data collection.

The Authors collected handwriting samples from approximately 300 writers. Every writer was a native Chinese speaker. Each writer was given a printed form showing the 20 most common Chinese words. Each writer was asked to write the 20 words 3 times yielding 60 individual exemplars per writer. The Authors’ initial findings, reported in this paper, focused on the six most simple character forms within this “top 20” selection. The remainder of this paper focuses on the findings obtained from this set. The six selected characters are *bu*, *da*, *ge*, *le*, *ren*, and *shang*. These characters are illustrated in Figure 8 below. Figure 8 contains a printed version, a handwritten exemplar similar to the printed version, and several other handwritten forms of the character. This figure shows the degree of variability and complexity which must be confronted when attempting recognition of Chinese words.

Numerous versions of these six characters were prepared for analysis by performing the following steps:

1. Scanning the handwritten documents into images.
2. Labeling all the words within the images to establish “ground truth”.
3. Converting the images into graphs.
4. Computing all embedded versions of these graphs.
5. Calculating the isomorphic graph keys for each full graph as well as each embedded graph.
6. Calculating the Feature Angles for each full graph plus all embedded graphs.
7. Creating a database of graph information.

Name of Character	Printed Version	Handwritten Exemplar	Alternative Handwritten Forms		
<i>bu</i>	不	不	不	不	不
<i>da</i>	大	大	大	大	大
<i>ge</i>	个	个	个	个	个
<i>le</i>	了	了	了	了	了
<i>ren</i>	人	人	人	人	人
<i>shang</i>	上	上	上	上	上

Figure 8: Selection of six commonly used Chinese words showing variations in written forms.

Before the testing methodology is fully described, two more points should be considered. The first is that the most reasonable exemplar for a particular written character may not be the embedding which covers 100% of the pixels of the character. This idea is represented in Figure 9. This figure contains a *ge* character. The 100% character in the left part of the figure contains what might be characterized as an “ink smear” which results in an isomorphism which is unnecessarily complex (8;112.24.0.0.32, an “H” with a line). The character on the right is the 94%

embedding which results in a simpler isomorphism more in keeping with many other *ges* (6;112.0.64, a “T” with a line). This example shows that taking subgraphs with embedding percents not equal to 100% can yield a more appropriate characterization of a writing.



Figure 9: A 100% (left) and 94% (right) embedding of a *ge*.

Another concept which is useful is the idea of an augmented isomorphism. When the lines of a writing contain extreme bends (such as an “S” shape, the augmented isomorphism adds a “soft vertex” in the character, so that more of the shape of the writing can be captured by the Feature Angles. An example of this is given in Figure 10. This figure shows the character *le* with original vertices in gray and the “soft vertex” in a white box containing an “x”. The presence of the soft vertex enables the Feature Angles to more closely represent the extreme bending of the line.

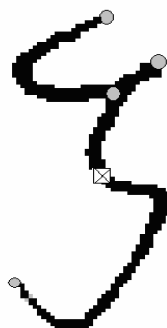


Figure 10: A *le* showing original vertices and augmented soft vertex.

The conjecture underlying the Authors’ analysis is that it is possible to identify empirically a set of embedded graph-forms that will perform the function of “radicals” in Chinese word recognition. The methods presented in this study represent a first step in a process of building a “bullet proof” reference collection of forms in support initially of word spotting and ultimately of word recognition. The first steps toward building such a reference set are discussed as follows.

The six selected Chinese words are all different in their structure, but they have multiple common embedded graphs of the same isomorphic class. The key to distilling embedded graphs that behave as “pseudo-radicals” is to find graphs that are unique in their geometry and encompass significant portions of

the overall Chinese word graph. The Authors’ selection of simple words was intended to test the concept of embedded pseudo-radicals while controlling for other issues of complexity occurring in Chinese words. Next steps will involve extending these concepts to more complex character forms.

The selected characters principally exhibited six common embedded isomorphic forms. These forms, with descriptive captions, are shown in Figure 11 as follows:

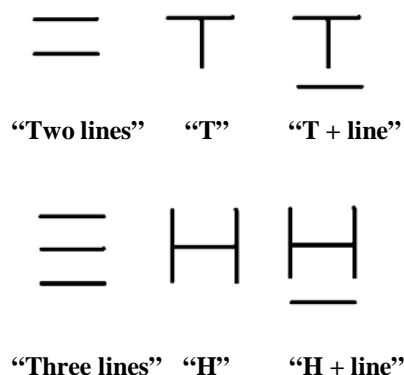


Figure 11: Six most common unique embedded graphs in six selected Chinese words.

Each of these graph types, extracted automatically, represents ones that occur both within a particular word and across different words at high embedding levels. It is the Authors’ belief that a similar set of graphs can be extracted for more complex Chinese words but these graphs will be observed at a lower embedding level.

The analysis entailed identifying all embedded forms within the six selected Chinese words. The percentage of embedding ranged from 100% (the full graph) to 60%. That is, the full graphs were decomposed into all graph forms encompassing 60% or greater of the pixel content of the graph. Each full graph as well as its embedded forms was classified in two ways: (1) *Topology*--through the isomorphic graph keys and (2) *Geometry*--through the Feature Angles. These two parameters represent a very effective and compact way to represent graphical forms at varying levels of complexity.

A study was conducted to determine whether using the augmented isomorphisms and Feature Angle measurements would enable an automated system to identify characters for use in word spotting. The data for a particular character were divided into a training set and a test set in proportion approximately two-thirds in the training set and one-third in the test set.

The training set was then used to determine a set of augmented isomorphisms and Feature Angles which describe the character under study. The test set was then compared to this set of augmented isomorphisms and Feature Angles to see if any matches were found. In addition *all* the writings for the other five characters were compared to the set of augmented isomorphisms and Feature Angles to see if any matches were found for these characters. Effectiveness was measured in two dimensions. The first is “recall” which represents the percentage of all the characters being sought in the test set which were actually identified. The second is “precision” and is the percent of all the characters which were matched by the process which were actually the character being sought. High values of both these measures are desirable. This entire process was replicated for different divisions into training and test data and for various combinations of three parameters which govern the process.

The three parameters are the following: The first is the tolerable deviation in an angle which could still be considered a match. This parameter is called delta. Larger values of delta make it easier to find a match and might be predicted to increase recall but possibly decrease precision. The second parameter is the lower limit for embedding level for a sub-graph in the training data to be eligible for the exemplar pool. The third parameter is the lower limit for embedding level for a sub-graph in the test data to be used as a test comparison. Larger values of these latter two parameters make it harder to find a match and might be expected to increase the precision and possibly to decrease the recall.

For word matching, recall and precision are not equal. High values of recall are more important, since the point of word matching is to save work for a human reader. Thus it is important to catch almost all the instances of the character sought. The human reader can use context and other cues to sort out the true instances from the false. Increased precision makes the task of the human reader easier, but 100% precision is not required for word matching to be effective.

The results for a particular experiment for two characters are given in Figures 12a through 12d. These figures give the mean results for three replicates for each combination of parameters. The values of delta used in this experiment were 12, 13, 14, and 15 degrees. The values of the model embedding cut were 87, 89, 91, 93, and 95 percent. The values of the test embedding cut were 81, 83, 85, and 87 percent. All combinations were run so the graphs in the figures are based on 240 runs each.

Figure 12a shows that the mean recall is over 80% for all combinations of the parameters. A characteristic “saw-tooth” pattern is evident. This pattern, which is present in Figure 12c as well, shows that increasing delta increases recall (other things being equal) while increasing the cut percentages decreases recall. The saw-tooth comes from the fact that resetting the test cut to 81% from 87% increases recall, but to a value somewhat smaller than that for the smaller value of the modeling cut. Figure 12c shows the same general pattern, but with greater decreases in recall associated with increases in the modeling cut.

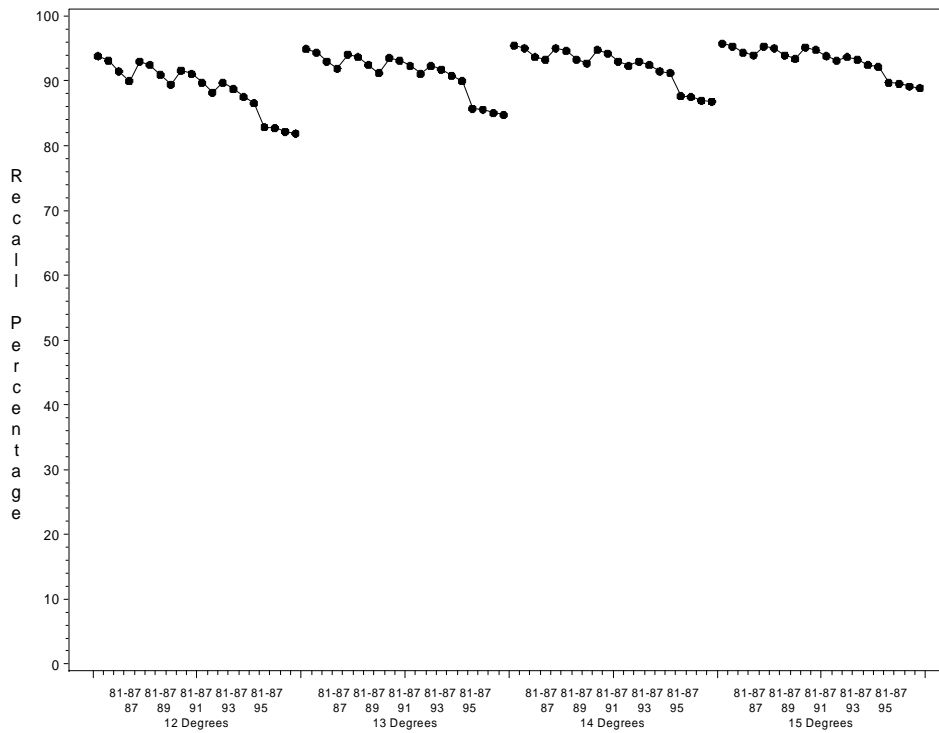
Figure 12b shows a severe saw-tooth pattern with precision generally decreasing with increasing delta and increasing with increasing values of the cut parameters. However, there is a large decrease in precision when the test embedding cut is changed back to 81% for the next higher level of the modeling cut. The saw-tooth pattern does not hold in Figure 12d. There is a general increase in precision with increased values of the cut parameters and decreased precision with increased delta.

To get a feel for what kind of performance could be expected from a “tuned” version of this process the following set of calculations was conducted. For each of five characters tested (*bu*, *ge*, *le*, *ren*, and *shang*), a search was conducted for the set of parameters which gave the highest value of precision for a minimum value of 92% for the recall. These results are presented in Table 1. These results show that it is hardest to achieve high precision for *ge* and *bu*, while much easier for *shang*, *le*, or *ren*.

Table 1: Best precision achievable for recall at least 92% for five characters

Character	Best Achievable Precision
<i>Bu</i>	45.7%
<i>Ge</i>	37.3%
<i>Le</i>	74.9%
<i>Ren</i>	78.0%
<i>Shang</i>	65.3%

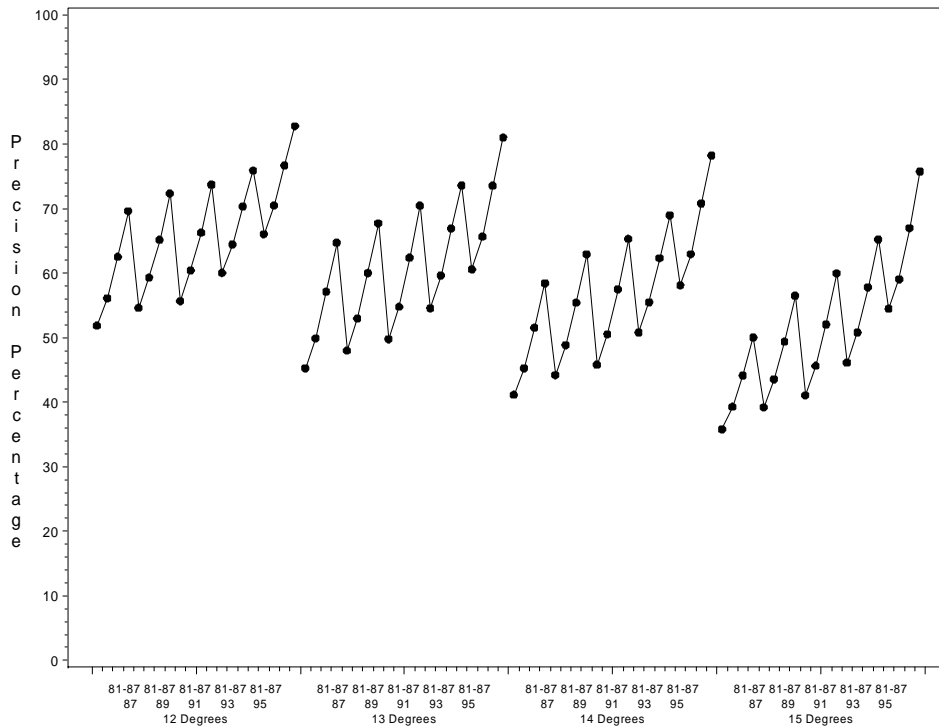
Recall Percentage for "Shang"



Parameters: Test Embedding / Model Embedding / Angle Delta

Figure 12a: Mean percentage recall for study of *shang*.

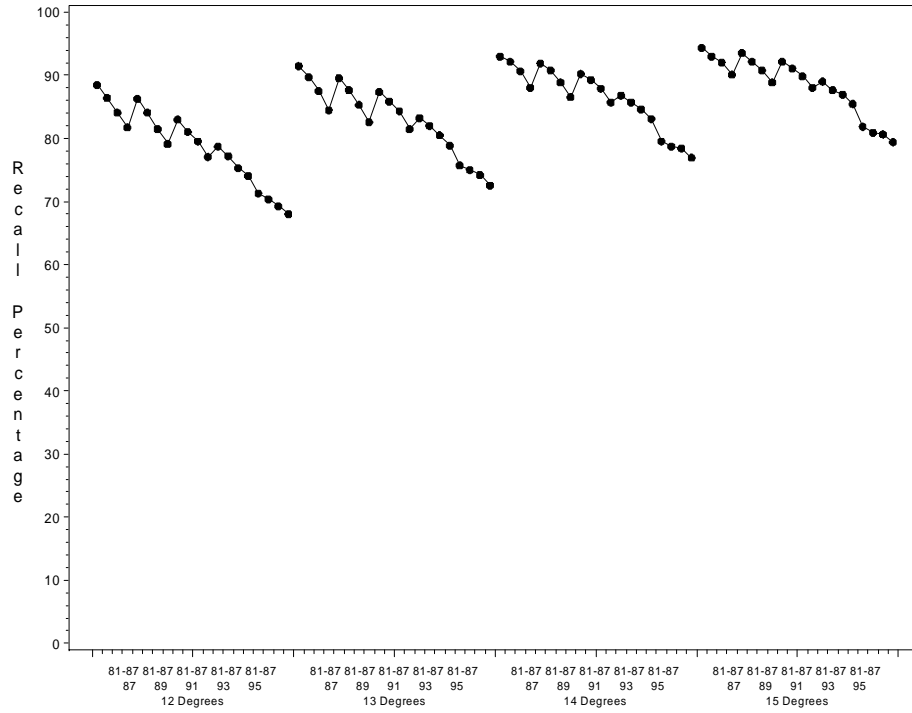
Precision Percentage for "Shang"



Parameters: Test Embedding / Model Embedding / Angle Delta

Figure 12b: Mean percentage precision for study of *shang*.

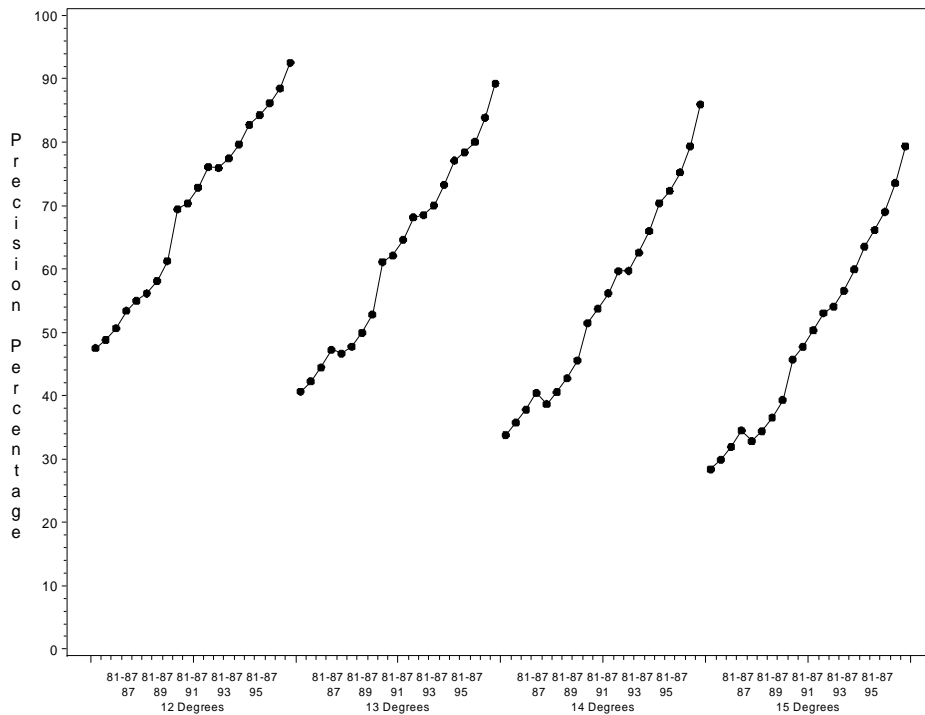
Recall Percentage for "Bu"



Parameters: Test Embedding / Model Embedding / Angle Delta

Figure 12c: Mean percentage recall for study of *bu*.

Precision Percentage for "Bu"



Parameters: Test Embedding / Model Embedding / Angle Delta

Figure 12d: Mean percentage precision for study of *bu*.

8 Conclusions

The work described in this paper represents the initial steps toward developing a viable means for Chinese word spotting (and ultimately recognition) based on Pictographic Recognition. The concept driving the Authors' approach was that it would be possible to empirically derive sets of embedded forms that characterized the Chinese word in which they were embedded. To test this idea, the Authors collected a sample of Chinese writings and focused on a subset of six simple character forms. These six forms were selected because they enabled the Authors to control for complexity during the initial evaluation. The six simple forms also presented challenges because of their similarity to each other and the high number of common embedded forms they shared.

The Authors were able to implement a system that automatically distilled embedded forms from handwritten Chinese words, compiled a database of these forms, and used these forms to recognize other versions of these same Chinese words. The factors that influenced the behavior of recognition based on embedded forms are:

1. Embedded graph topology as expressed through a code corresponding to graph isomorphism.
2. Embedded graph shape as described through Feature Angles relative to a designated node within the graph.
3. Percentage of embedding as described as the proportion of pixels from the entire Chinese word occupied by the embedded graph.

In the system implemented by the Authors, embedded graphs were shown to behave in a manner similar to the actual building blocks of written Chinese: the radicals. These pseudo-radicals empirically generated within this study confirm the concept that is possible to isolate common topological and geometric forms that transcend multiple occurrences of the same written Chinese word. And, once isolated these common embedded forms can become the foundation for word recognition.

The Authors intend to continue this work by focusing on the following topics:

1. Continuing to build a ground truth database of Chinese handwriting exemplars.
2. Applying the techniques herein described to more complex Chinese words.

3. Continuing to “tune” the recognition process based on the three parameters discussed above: Topology, Geometry and Embedding.
4. Expanding from the concept of Feature Angles to more detailed measurement data. It should be noted that Pictographic Recognition offers a wealth of feature information that was intentionally excluded from the present study that will be brought to bear on future work.
5. Applying more sophisticated scoring methods—currently used for English and Arabic recognition to Chinese. These methods permit composite scoring among numerous embeddings using a “preponderance of evidence” approach.