Statistical Language Models for Graphical Object Recognition

Laura Keyes[1], Andrew O'Sullivan[1] and Adam Winstanley[2]

[1] School of Informatics and Engineering, Institute of Technology Blanchardstown, Dublin 15
[2] Department of Computer Science, NUI Maynooth, Maynooth, Co. Kildare
Contact email: Andrew.O'Sullivan@itb.ie

Abstract.

This paper explores automatic recognition and semantic capture in vector graphics for graphical information systems. The low-level graphical content of graphical documents, such as a map or architectural drawing, are often captured manually and the encoding of the semantic content seen as an extension of this. The large quantity of new and archived graphical data available on paper makes automatic structuring of such graphical data desirable. A successful method for recognising text data uses statistical language models. This work will investigate and evaluate similar and adapted statistical models (Statistical Graphical Langauge Models, SGLM) to graphical languages based on the associations between different classes of object in a drawing to automate the structuring and recognition of graphical data.

Keywords *Statistical Language Models, Semantic Modelling, CAD Drawings, Graphical Object Recognition, Statistical Graphical Langauge Models, Operation and Maintenance Information System*

1. Introduction

Graphical information systems are computerised systems used for storing, representing, manipulating, analysing and displaying graphical data. The increased used of graphical information systems has motivated research in the automatic structuring of graphical data and developing and applying graphical object recognition. That is, a vast amount of data archived by organisations is in graphical form (for example, diagrams, maps, technical drawings, and architectural plans). For this to be searched, analysed and synthesised automatically, it must be parsed and converted from simple graphics (points, lines, symbols, polygons) to semantically rich graphical information ("circuit-breaker", "building", "spark-plug", "extractor fan"). For computer systems to process such graphical data not only the geometry but also attribute data, describing the nature of the objects depicted must be stored.

This manual structuring into composite objects and addition of labelling attributes is a labour-intensive, expensive and error prone process. The successful automation of raster-vector conversion plus the large quantity of new and archived graphical data available on paper makes the automation of feature extraction and structuring of graphical data desirable. Automation of the structuring and recognition of objects through statistical modelling for efficient and complete input into graphical information systems can form a solution to this complex problem.

Statistical language models are a successful method for recognising text data. These models are derived from corpora of language-examples using the frequency and associations between words. This work will apply and evaluate similar and adapted statistical models *(Statistical Graphical Langauge Models, SGLM)* to graphical languages based on the associations between different classes of object in a drawing to automate the structuring and recognition of graphical data.

This paper describes the proposed research into the use and adaptation of SLM techniques to aid in the semantic analysis of graphical data for the purposes of recognition, indexing and retrieval. The derived graphical recognition system will be used for the development of an operation and maintenance information system for architectural plans within buildings and other facilities (Entropic Ltd)[*].

2. Operation and Maintenance Information System

An Operation and Maintenance (O&M) information system holds centrally all relevant information pertaining to the operation and maintenance of plant and equipment within buildings and other facilities. This information is presented through a multi-media web interface and consists of drawings, data sheets, operating instructions, parts listings, suppliers, installers, manufacturers and other details of all the service utilities. The information on each component is comprehensively cross-referenced using links between corresponding items in drawings, data sheets, photographs and so on. The system can be implemented for all sizes of installations but comes particularly suited for the infrastructure

[*] Entropic Ltd, are a SME located in County Kildare, Ireland and are exploring the provision of multimedia operation and maintenance information systems for building and plant facilities management.

management of large industrial or service sites. Current use includes a sports complex and large private dwellings.

The Operation and Maintenance Information System allows a user to select an example object (simple or composite) and the software finds similar objects in the same or other drawings. The tool generates data structures that can be used to build multimedia linkages between objects, drawings and related information. The information is accessed through a standard web browser interface including navigation through hot-links and key-word search facilities. CAD drawings showing the location of utilities and services also act as browser navigational maps. In operation, the system's main use concerns day-to-day operation and maintenance tasks**,** for example**:**

- Retrieving plant operating and servicing instructions
- Scheduling of maintenance tasks
- Keeping records of maintenance done
- Listing of spare parts
- Locating rarely accessed equipment, plant and components
- Generating service reports

2.1. Problems of Data Capture and Construction

A typical O&M system has to be compiled from information supplied by many manufacturers, architects, designers and contractors in a wide variety of formats: CAD drawings, data sheets, operating instructions, parts listings, details of suppliers, installers and manufacturers. Some are available digitally but many are paper documents. O&M systems commissioned so far have been constructed manually through digitising, structuring and linking this information appropriately.

For the system to be economic, it is desirable to automate as much as possible of this compilation process. Automation possibilities include:

- Recognition and labelling objects/components on drawings
- Generating links through string matching
- Compilation of databases of information from scanned text/drawings

Once recognised and classified, these objects can be assigned unique identifiers in the system. This allows their inclusion in the search and navigation functions. Previous work evaluated the recognition and labelling of objects and components and drawings using shape [11] and structural descriptors [14]. As part of this project, for the automatic structuring and

recognition of technical data for a web-based multimedia O&M information system, an adapted SLM technique will be used. This work will also investigate if SGLM can be applied to improve recognition performance of shape and structural methods to provide an optimal solution to the problem of graphics recognition for architectural and engineering graphical domains.

3. Graphical Object Recognition and SLMs

Graphics recognition involves the recognition and structuring of geometry such as points, lines, text, symbols on graphical documents into meaningful objects for use in graphical information systems. Graphics recognition is a sub-field of pattern recognition and includes classification and recognition of graphical data based on shape description of primitive components, structure matching of composite objects and semantic analysis of whole documents. A sub-field of semantic analysis is to treat the graphical notation as analogous to textual language by, for example, constructing a graphics parser based on a formally defined grammar.

Statistical language models have been used with natural language processing applications such as speech recognition and spoken language understanding. They are based on the analysis of a large corpus of text to construct a probabilistic contextual model for the occurrence of words (and/or larger structures). The model is used to increase the effectiveness of other recognisers.

This work will investigate the use and adaptation of SLM techniques to aid in the semantic analysis of graphical data for the purposes of recognition, indexing and retrieval. A number of techniques (n-gram models, hidden Markov models, part-of-speech tagging) will be adapted and evaluated for graphical data. A rational for their use will be formulated. A categorisation of the different domains of graphical data by form and content will be made. Software modules will be created to test and illustrate Statistical Graphical Language Model (SGLM) techniques' effectiveness on the architecture and engineering domain.

The suggestion that this may be a valid approach is re-enforced by the similarities between textual and graphical notations [1]:

- Both consist of discrete objects (words, graphical objects)
- Objects have a physical form (spelling/pronunciation, shape)
- Objects have a semantic component (meaning, graphical object label)

- Objects are classified according to function (part of speech, object class)
- Objects are formed into larger components (sentences/paragraphs etc., regions/diagrams etc.).

Depending on the nature of the graphical notation, this analogy can be very strong. For example, at one extreme, visual programming languages have precise grammars that can be used to create well-formed software tools to edit, check and translate valid programs. Other notations, while containing conventional symbols, are depictions of the real-world configuration of objects that has a much less structured syntax, although there is usually some underlying structure. For example [18], on a map a building needs access to a road that has connections to other roads, and so on. Part of the proposed research is to characterise the applicability of SLMs to each subject domain according to this underlying structure. Of course, there are differences between natural language and graphical notations:

- Natural language is one-dimensional; graphics are usually two-dimensional.
- Natural language is sequential - the meanings of sentences are determined by the order of their component words; graphical notations use more complex spatial relationships.
- The vocabularies in natural language texts are generally larger than the symbol vocabulary of most graphical notations.

The proposed research will assess how these differences affect the applicability of SLMs and how they can be incorporated into a SGLM. Also, SLMs will be investigated and evaluated on the problem of automatically recognising and interpreting graphical data on technical drawings for the development of an operation and maintenance information system for plans within buildings and other facilities.

3.1 Statistical Language Models

Statistical Language Models are estimates of probability distributions over natural language phenomena such as sequences of letters, words, sentences or whole documents. They were first used by Andrei A. Markov at the beginning of the 20th century to model letter sequences in Russian literature [13]. While this was a linguistic task, these methods were then developed as a general statistical tool. They have been primarily developed for natural language processing. Automatic speech recognition is arguably the area that has benefited the most from SLMs where they have proved quite successful [7]. A possible system architecture (to improve speech recognition) is shown in figure 1. SLMs have also been used in the fields

of machine translation, optical character recognition, handwriting recognition, information retrieval, augmentative communication systems and many more [8].

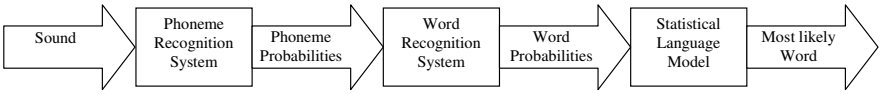

Figure 20 Typical speech recogniser.

SLMs employ statistical estimation methods that make use of large corpuses of training data in the form of text. These corpuses can consist of thousands or millions of words from a language. In order to be as representative as possible of a language, a corpus usually has text from a wide variety of sources. For example, the derived *Brown Corpus* [13] consists of one million words taken from fifteen different categories such as legal text, scientific text and press reportage. A corpus can however be built to just include a particular sub-set of language, if so required for a particular task. Generally the larger the corpus the better it will be for statistical language modelling.

3.1.1 N-gram models for SLM

A SLM is simply a probability distribution *P(s)* over a sequence of words (or sentences or whole documents and so on). In practice it is impossible to know the probability so instead the estimate of the probability is used. This estimate is found by using the frequency of text within the training data. Generally a language model is represented as a conditional probability distribution of the next words to be seen, given the previous words, that is:

$$P(w_i \mid h_i), \text{ where } h_i = (w_1, w_2, ..., w_{i-1}) \text{ and } w_i \text{ is the } i^{\text{th}} \text{ word} \quad (1)$$

The purpose of a SLM is to assign high probabilities to likely word sequences and low probabilities to unlikely ones. Different SLM models can be combined using techniques such as linear interpolation. N-gram models are the most widely used SLM technique. They use the previous *n*-1 words to predict the next word. Generally *n* is either 2 (a bi-gram), 3 (a tri-gram) or 4 (a four-gram). A bi-gram model is looking for the probability $P(w_i \mid w_{i-1})$ and a tri-gram model is looking for the probability $P(w_i \mid w_{i-1}, w_{i-2})$. These probabilities are estimated by using relative frequency:

$$P(w_i \mid w_{i-1}) = C(w_{i-1} w_i) / C(w_{i-1}) \quad (2)$$

and

$$P(w_i \mid w_{i-1}, w_{i-2}) = C(w_{i-2}, w_{i-1} w_i) / C(w_{i-1}, w_i) \quad (3)$$

where *C* is the frequency of the enclosed words in the training corpus. For example if a sentence starts with "I was walking the" a tri-gram model would use the two words "walking the" to predict the next word. This prediction is done using the training data corpus. The corpus is analysed for co-occurrences of words, in this case triples that start with "walking the". The triples are sorted in terms of the frequency they appear in the training data, with the most frequent triple the one used for the prediction. To use this example, the training data may have the triple "walking the dog" as the most frequent triple that starts with "walking the" so the word "dog" is given as the prediction.

There are other SLM techniques which are also used [15]. These include *Decision Tree* models [2] which assign probabilities to each of a number of choices based on the context of decisions. Some SLM techniques are derived from grammars commonly used by linguists. For example Sjilman et al. [16] uses a declarative grammar to generate a language model in order to recognise hand-sketched digital ink. Other methods include *Exponential* models and *Adaptive* models. [15] suggests that some other SLM models such as *Dependency* models, *Dimensionality* reduction and *Whole Sentence* models show significant promise. However this research will focus on the most powerful of these models the N-gram and its variants [18].

There are problems that affect SLMs. One problem is the data sparseness problem. This problem is simply that a training corpus, no matter how big cannot cover all probabilities. These probabilities are then automatically assigned a zero value. So when a phrase occurs that has not been seen before, that is, it is not in the training data, its probability is zero. To solve this problem techniques are used that assign a 'non-zero probability' to 'zero probability'. This process is called *Smoothing* [13,7,8].

3.1.2 Evaluation of SLM

In order to compare SLMs common measures used. These are based on the concepts of relative entropy, cross entropy and perplexity [13]. Combined with the use of standard corpora and test data sets, they provide for the calculation of objective metrics for SLMs.

Entropy is a measure of information in a random variable. It can be used as a metric to measure how much information there is in a particular grammar, and also to measure how well a given N-gram model will be able to predict the next object. Computing entropy requires that we establish a random variable *X* that ranges over a sequence of objects (the set

of which we will call χ) and that has a particular probability function, call it $p(x)$ the entropy of this random variable X is then

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \quad (4)$$

Entropy is measured in bits. The lower amount of the entropy we get the best model we have. The value of 2^H is the perplexity. Perplexity can be thought of as the weighted average number of choices a random variable has to make [8]. It can be seen as a measure of the size of the set of words from which the next word is chosen from. Generally the lower the perplexity the better the model.

3.2 Applying SLMs to Graphics

The success of statistical language models has been due to the efficiency of these models and to the linear structure of natural language utterances and the underlying grammar (the semantic and syntactic relationships between adjacent words). In graphical data, there is no rigid grammatical structure. However, a quasi-grammatical pattern does exist (for example, vent-duct-fan or witch-wire-socket) and this suggests that the language model approach may have some validity. However, unlike natural language, these sequences have no inherent direction.

Given the similarities between graphics and natural language, it seems reasonable that SLMs may have applicability to improve the classification of graphic objects as they do for natural language processing applications. On major difference is that, whereas language is naturally a one- dimensional sequence of symbols, graphics are inherently multiple-dimensional. Therefore, for direct application, it is necessary to extract one-dimensional sequence from the graphical data. One approach of doing that is to use adjacency relationships between objects on a drawing/document. Alternatively, the SLM theory can be extended to deal with two-dimensional “sequences”.

Within this work SLMs will be used to measure the frequency of each graphical objects context allowing a graphics recognition system to be constructed in a similar way used for a speech recognition system (figure 2). In figure 2 the system depicted would be used to extend the classification capabilities of other recognition methods for example, based on an object's shape. The image is vectorised, cleaned and topologically corrected to form polygons. A

recognition system produces probabilities for candidate classes of each object based in this case on their shape [11]. The SLM, built from analysis of another data set, uses the probabilities to construct "phrases" of objects. A shape recognition system produces probabilities for the candidate classes of each object. The statistical language model uses these probabilities to construct candidate "phrases" of objects and use the n-gram model built from a corpus to select the most likely candidate object class.

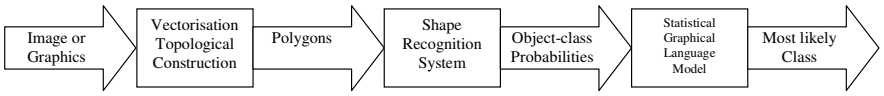

Figure 21 Possible graphical object recognition system with SGLM (see figure 1)

4. Graphical Recognition System

A main outcome of this work will be a software module that can be used and evaluated in the production process of O&M systems. Figure 3 shows the software configuration envisaged and the role of SGLM within this system. Digitised CAD drawings of the building/plant services will be processed to extract their component objects from which shape and structural descriptions are built. These feed into several description and matching algorithms, each of which produces one or more candidate categories to which each object may belong. A fusion algorithm produces an overall consensus decision giving a ranked list of candidate types. The SGLM module can then be used to improve the performance of the recognisers.

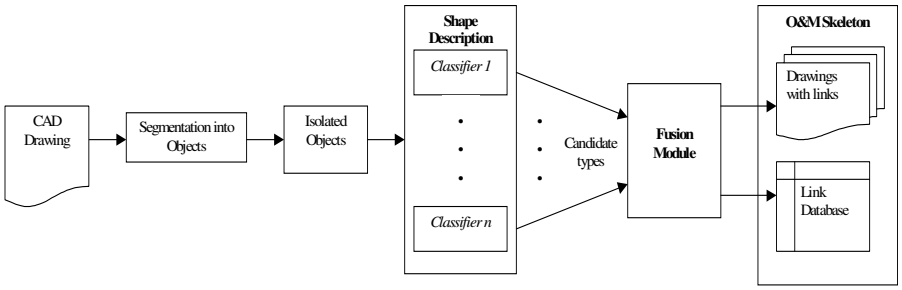

Figure 22 Graphical Shape Recognition System configuration

4.1 Evaluating the SGLM system

To evaluate the classification performance, *precision*, *recall* and *accuracy* (defined below) will be used. These notations are frequently used in information retrieval (IR) applications to evaluate statistical NLP models, and their use has crossed over into work on evaluating SLMs for many problems. Precision is defined as a measure of selected objects that the classification system got right.

$$precision = \frac{tp}{tp + fp} \quad (5)$$

Where *tp* (true positive) and *tn* (true negative) account for the cases the classification system got right and the wrongly selected cases in *fp* are called false positive. The cases in *fn* that failed to be selected are called false negative.

Recall is defined as, the proportion of the target objects that the system selected.

$$Recall = \frac{tp}{tp + fn} \quad (6)$$

Accuracy is defined as, the proportion of correctly classified objects.

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn} \quad (7)$$

Fallout is a less frequently used measure. It is defined as the proportion of non-targeted items that were mistakenly selected and is defined as follows:

$$fallout = \frac{fp}{fp + tn} \quad (8)$$

Intense evaluation of the system forms part of the overall research goal.

5. Conclusion and Future Work

Treating graphical notations as examples of language is well established and the use of syntactic grammars to generate or parse graphical is well known. Similarly, the development of statistical natural language models is advanced. However, the aim of this work is the application of statistical language models to graphical notations. By identifying graphical notations properties that make them suitable for these models, this research will offer a theoretical foundation for new methods of capturing, searching and analysing graphical data.

This work has relevance to sectors that collect, supply or use graphical data in digital form. There are enormous amounts of data in paper form, examples come from surveying, mapping, architecture, engineering and multimedia systems. Aside from the architectural and engineering domains identified for use in this work, it is envisaged that this research will result in software modules that can be used in various configurations for different application domains. For example, recognition and retrieval of graphical data for multimedia operations, automatically structuring geometry, detection and correction of errors in structure for graphics recognition.

Acknowledgements

This project is supported by Technological Sector Research: Strand 1 Post-Graduate R&D skills programme and Entropic Ltd, Maynooth, County Kildare.

References

[1] Andrews, J.H., Maps and language, A metaphor extended, *Cartographic Journal*, 27, 1–19, 1990.

[2] Bahl, L R., Brown, P. F., Peter V. de Souza and R. L. Mercer., A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37:1001-1008, July 1989.

[3] Baum, L.S. et al., Extracting System-level Understanding from Wiring Diagram Manuals, *5th IAPR International Workshop on Graphics Recognition* (GREC 2003), 132-138, Barcelona, July 2003.

[4] Brown, P.F. et al., A Statistical approach to machine translation, *Computational Linguistics*, 16 (2), 79–85, 1990.

[5] Cutting, D., Kupiec, J., Pedersen, J. and Sibun, P., A Practical Part-of-speech Tagger, *Third Conference on Applied Natural Language Processing*, (ANLP-3), 133-140, 1991.

[6] Delandre, M., Trupin, E., and Ogier, J-M., Local Structural Analysis: a primer, *5th IAPR International Workshop on Graphics Recognition* (GREC 2003), 277-285, Barcelona, July 2003.

[7] Jelinek, F., Statistical Methods for Speech Recognition. MIT Press 1997.

[8] Jurafsky, D. and .Martin, J.H., Speech and Language Processing, Prentice-Hall, 2000.

[9] Keyes, L. and Winstanley A.C., Automatically Structuring Archaeological Features on Topographic Maps. *GIS Research UK*, 191-4, Sheffield, April 2002.

[10] Keyes, L., Winstanley, A., and Healy, P., Comparing Learning Strategies for Topographic Object Classification, *IEEE International Geoscience and Remote Sensing Symposium* (IGARSS'03), July 2003.

[11] Keyes, L., A. Winstanley: Shape Description for Automatically Structuring Graphical Data, *Graphics Recognition - Recent Advances and Perspectives*, Series: Lecture Notes in Computer Science, Vol. 3088 J. Lladós and Y.B. Kwon (Eds), Springer-Verlag, August 2004.

[12] Kittler, J., Hatef, M., Duin, R.P.W and Matas, J., On Combing Classifiers, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 20 (3), 226-239, 1998.

[13] Manning, C.,D., and Schutz, H., Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, 2001.

[14] O'Donoghue, D., Winstanley, A., Mulhare, L. and Keyes, L., Applications of Cartographic Structure Matching, *IEEE International Geoscience and Remote Sensing Symposium* (IGARSS'03), July 2003.

[15] Rosenfeld, R., Two Decades of Statistical Language Modeling: Where Do We Go From Here?, *Proceedings of the IEEE*, 88 (8), pp 1270-1278, 2000.

[16] Shilman, M., Pasula, H., Russell, S. and Newton, R., Statistical Visual Language Models for Ink Parsing. *AAAI Spring 2002 Symposium on Sketch Understanding*, 2002.

[17] Winstanley, A.C. and Keyes, L., Applying Computer Vision Techniques to Topographic Objects, *Int.l Archives of Photogrammetry and Remote Sensing*, 33 (B3): 480-487, (2000).

[18] Winstanley, A.C., Salaik, B., and Keyes, L., Statistical Language Models For Topographic Data Recognition, *IEEE International Geoscience and Remote Sensing Symposium* (IGARSS'03), July 2003.