TECHNOLOGICAL SECTOR RESEARCH: STRAND I

POST-GRADUATE R&D SKILLS PROGRAMME

2004 Application Form

Notes:

1. The General Guidelines and Conditions on the Post-Graduate R&D Skills Programme **must be read** prior to completing this application form. **Applications not adhering to the Guidelines will be returned to the Institute concerned.**

2. In particular, please note that completed applications must **not exceed 20 pages** in length (excluding appendices). **Applications over 20 pages in length will be returned to the Institute concerned.**

3. Appendices consist of appropriate letters of support and pro-forma supervisor profile sheets **ONLY**.

4. Please have regard for the Marking Scheme prior to completing this Application Form.

5. Please ensure that you complete **all** sections of this form as fully as possible in typescript.

6. All Sections must be typed in at least **12pt font**.

. 5. The Abstract on Page 2 of this form must be completed. The abstract should set out in **non-technical language** the nature of the proposed research project and what it is hoped will be achieved, i.e., the likely application of research outcomes. **Please note** that this does not obviate the need, where required, for appropriate technical description on the application form. The abstract **must not exceed 200 words**.

FOR OFFICE USE ONLY		
Project Code PRDSP/03/	**Institute**	**Category**

Institute:	Institute of Technology Blanchardstown			
Project Title:	Statistical Language Modelling for Graphical Object Recognition			
Category[1]: (insert appropriate category code from list)		IT		
Lead Supervisor:	Laura Keyes			
Has this project been submitted to Strand I before?		No	Outcome of this submission?	
No. of ongoing Strand I projects with lead supervisor		0	No. of completed Strand I projects with lead supervisor	0

Please summarise below the proposed project, the potential research outcomes and their likely application in <u>non-technical language</u>. This abstract should <u>not exceed 200 words</u>.

ABSTRACT

A vast amount of data archived by organisations is in graphical form (for example, diagrams, maps, technical drawings, 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"). Usually, this structuring into composite objects and the addition of labelling attributes is a manual, labour-intensive, expensive and error-prone process. Previous work has evaluated the recognition of objects on drawings and plans using shape and structural descriptors and although successful (75%-80% recognition rate) they are not an optimal solution to the problem. A successful method for recognising text data uses s*tatistical language models*. These models are derived from corpora of language-examples using the frequency and associations between words. It is proposed to apply similar and adapted statistical models *(Statistical Graphical Langauge Models, SGLM)* to graphical languages based on the associations between different classes of shape in a drawing to automate the structuring and recognition of graphical data. This work will be conducted in collaboration with a university and also an industrial partner dealing with a web-based operation and maintenance information system for buildings and industrial premises.

Lead Project Supervisor/Department:	**Laura Keyes**				
	School of Informatics and Engineering				
Institute:	**Institute of Technology Blanchardstown**				
Telephone:	**01-8851339**	**Fax:**	**01-8851001**	**Email:**	**Laura.Keyes@itb.ie**

Second Project Supervisor/Department:	[redacted]				
[redacted]	[redacted]				
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

1. Costs in Respect of Project

[redacted]			
[redacted]			
[redacted]			
[redacted]		[redacted]	

[redacted]

[redacted]

2 General Outline of Project Proposal

Main Research Question(s):

- Statistical Language Models are well known models and techniques used for natural language processing on textual data. This proposal is to investigate if SLMs can be adapted to deal with graphical notations i.e. Statistical Graphical Language Models (SGLM).
- Graphics recognition is a large active research area and uses many techniques and methodologies, including shape and structural and semantic analysis. This work will evaluate the effectiveness of a derived SGLM methodology in 'real-world' graphical object recognition for multimedia systems in architectural and engineering domains.
- Existing recognition systems based on shape and structural descriptors used for graphics recognition prove successful with 75% - 80% classification confidence. This work will assess the applicability and integration of SGLMs with existing work to improve classification performance and provide an optimal solution to the problem of graphics recognition.

Objectives:

- Investigate rigorously the theory of SGLMs
- Develop standard copora, test datasets and metrics
- Assess their applicability for architectural and engineering graphical domains
- Integrate them with existing work on shape and structural descriptors
- Build a graphics recognition software system based on SGLM/shape/structural methodologies
- Present and publish work at national and international conferences and seminars

Contribution to Knowledge of Discipline Area:

Graphics recognition is a sub-field of pattern recognition and includes classification and recognition of many types of graphical data (e.g. maps, architectural plans) based on shape description (e.g. Fourier Descriptors) 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.

A coherent body of theoretical knowledge and practical tools have been developed in both the graphics recognition and statistical language models fields. 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 novel aspect of the proposed research 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 provide the theoretical foundation for new methods of capturing, searching and analysing graphical data.

Relevance to Research Strategy of Institute:
The Institute of Technology Blanchardstown, although a newly established Institute, is actively partaking in and demonstrating its commitment to research and development through instigating several funded research projects and activities such as ITB conference, now hosted annually. Research and Development is a core activity at the Institute and in order to provide necessary skills and training of graduates, ITB seeks to actively engage with industry. ITB see collaborative industry-oriented research as a key component of its future strategy. This research is being supported by Entropic Ltd, who are a SME located in Maynooth, County Kildare and are exploring the provision of multimedia Operation and Maintenance information systems for building and plant facilities management and as such is directly relevant to the research strategy of the Institute.

3. DETAILED PROJECT DESCRIPTION

(To include justification of costs where necessary)

1. Project background

A division of Entropic Ltd. deals with the provision of multimedia Operation and Maintenance (O&M) information systems for building and plant facilities management. The 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 management of large industrial or service sites. Current use includes a sports complex and large private dwellings. Current negotiations are with an airport maintenance facility and a large multinational company.

The information is accessed through a standard web browser interface including navigation through hot-links and key-word search facilities. The 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

Data Capture for O&M

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 and also the quality assurance tests of the final system. Automation possibilities include:

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

Previous work has evaluated the recognition and labeling of objects and components on drawings and plans based on their shape [1, 14]. Shape is an important part of the semantic content of an object within a graphical information system. Shape description methods used include Fourier Descriptors (FD) and Moment Invariants (MI) and Scalar Descriptors (for example, area, elongation, number of corners etc.). These techniques are applied to object boundaries extracted from drawings represented as vector descriptions. The output obtained by the description methods provides a measurement of shape that characterises the object type. Each shape description method has proved successful in distinguishing graphical objects, with classification confidence of 75%-80%, however, no one technique provides an optimal solution to the problem.

This proposal is to investigate and perform recognition and classification of this graphical data using an adaptation of SLM [8,9] techniques. 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.

To this end, the proposed research will include a working liaison with Entropic Ltd. through:

- the supply of sample and test data for use in the project
- regular meetings with the participants to review progress and advise on the industrial and commercial aspects and
- availability for consultation on the project at other times.

2. Description of the proposed research

Previous research work carried out by the proposers in graphical object recognition used shape [1] and structural descriptors [12]. Practical applications have also been developed in the fields of large-scale cartography [14] and architectural plans [1]. Initial work in the proposed project will involve the re-engineering of these software modules (providing unsupervised and supervised learning capabilities) and the preparation and compilation of data for the domains being explored (architectural and engineering). Preparation and compilation of the data will include scanning plans/documents, vectorisation of the data, cleaning of the data to make it topologically sound and object extraction and segmentation (closed polygons, used by the recognition techniques).

It is proposed to 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 [4]:

- 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 [11], 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.

2.1 Traditional Statistical Language Modelling

Statistical language modelling is the attempt to capture regularities of natural language for the purpose of improving the performance of various language processing applications (e.g. speech recognisers, speech synthesisers, optical character recognisers, automatic translation systems, information retrieval and document analysis systems). In general, SLMs consist of estimating the probability distribution of various linguistic units such as words, sentences and whole documents and using this to predict the next unit in a sequence. A possible system architecture (to improve speech recognition) is shown in figure 1.

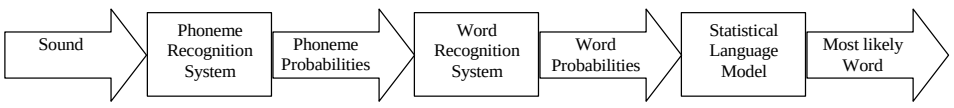

Figure 1: Typical speech recogniser.

SLM employs statistical estimation techniques using language training data in the form of text. Due to the categorical nature of language, and the large vocabularies people naturally use, statistical techniques must estimate a large number of parameters, and consequently depend critically on the availability of large corpora.

N-gram models for SLM

A statistical language model is a probability distribution over a sequence of words. A language model is generally represented as a conditional probability distribution of the next words to be seen, given the previous words, that is

$$P(w_i \mid h_{i-1}), \text{ where } h_{i-1} = w_1, w_2, ...w_{i-1} \quad (1)$$

Different models can be constructed depending on the length of the sequence of words used, for example unigram (*i = 1*), bigram (*i = 2*) and trigram (*i = 3*). These can be combined using linear interpolation, for example:

$$p(w \mid h_i) = \lambda_1 p_1(w \mid h_i) + \lambda_2 p_2(w \mid h_i) + \lambda_3 p_3(w \mid h_i) \quad (2)$$

where $\lambda_1 + \lambda_2 + \lambda_3 = 1,\ h_i \geq 0$.

The purpose of SLM is to assign high probabilities to likely word sequences and low probabilities to unlikely word sequences.

There are many types of statistical language models, however this research will focus on the most powerful of these models the N-gram and its variants [11].

N-gram models are typically constructed from statistics obtained from a large corpus of text using the co- occurrence of words to determine word sequence probabilities. These models have the advantage of being able to cover a much larger language than would normally be derived from a corpus. In contrast to other statistical language models, such as the grammar language model, n-gram models rely on the likelihood of sequence words such as word pairs (in case of bigram) or word triples (in the case of trigram) therefore they are less restrictive. Open vocabulary applications are easily supported with n- gram models. The use of stochastic n-gram models has a long and successful history in research community and is now more and more affecting commercial systems.

2.2 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 [1]. 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.

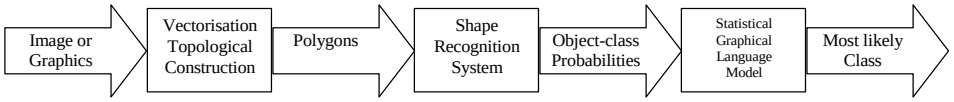

Figure 2: Possible graphical object recognition system (see figure 1).

2.3 Evaluation of the SGLM system

There is a large body of work from information theory that allow us to measure and compare the effectiveness of different methods of classification. These are based on the concepts of relative entropy, cross entropy and perplexity [8]. 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

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{3}$$

Entropy is measured in bits. The lower amount of the entropy we get the best model we have. The value of 2^H is called the perplexity. Perplexity can be viewed as the weighted average number of the choices a random variable has to make.

To evaluate the classification performance, we will adopt notations, such as precision, recall and accuracy (defined below). 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.

$$precision = \frac{tp}{tp + fp} \tag{4}$$

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.

$$Recall = \frac{tp}{tp + fn} \tag{5}$$

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

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn} \tag{6}$$

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:

$$fallout = \frac{fp}{fp + tn} \tag{7}$$

Intense evaluation of the system will form part of the overall project goal.

2.4 Graphics recognition system

The proposed project incorporates many research components, which need to be investigated and analysed. These include:

- investigate rigorously the theory of various SGLMs,
- develop standard copora, test data sets and metrics
- assess their applicability for architectural and engineering graphical domain and
- integrate them with existing work on shape and structural descriptors

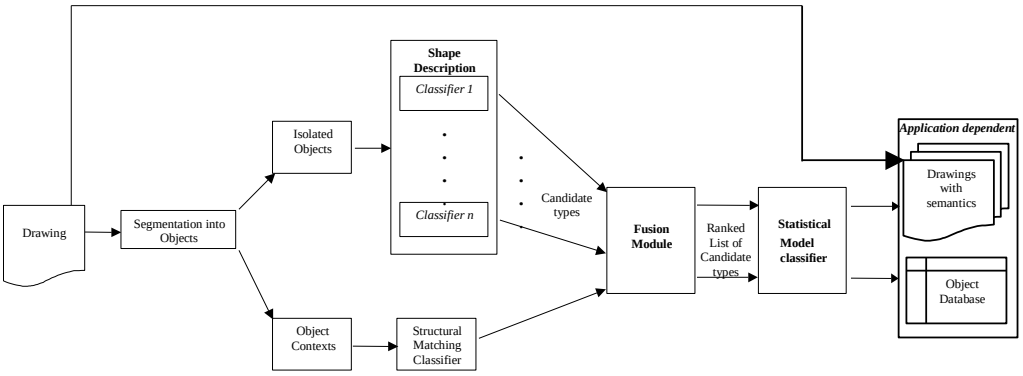

Figure 3. Proposed integration with existing classifier modules.

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.

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 novel aspect of the proposed research 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 provide the theoretical foundation for new methods of capturing, searching and analysing graphical data.

This work has relevance to industrial 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 project, 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.

Justification of costs

Materials:

[redacted]

Travel:

[redacted]

References

[1] Keyes, L., A. Winstanley: Shape Description for Automatically Structuring Graphical Data, *5th IAPR International Workshop on Graphics Recognition* (GREC 2003), 318-328, Barcelona, July 2003.

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

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

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

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

[7] L. Keyes and A.C. Winstanley "Fourier Descriptors as a General Classification Tool for Topographic Shapes", *IMVIP'99 Proceedings of the Irish Machine Vision and Image Processing Conference*, 193-203, Dublin City University, 1990.

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

[9] D.Jurafsky and J.H.Martin, *Speech and Language Processing*, Prentice-Hall, 2000.
[10] P.F.Brown, et al. "A Statistical approach to machine translation" *Computational Linguistics*, 16 (2), 79–85, 1990.
[11] Winstanley, A., B. Salaik, L. Keyes: Statistical Language Models For Topographic Data Recognition, *IEEE International Geoscience and Remote Sensing Symposium* (IGARSS'03), July 2003.
[12] O'Donoghue, D., A. Winstanley, L. Mulhare, L. Keyes: Applications of Cartographic Structure Matching, *IEEE International Geoscience and Remote Sensing Symposium* (IGARSS'03), July 2003.
[13] Keyes, L., A. Winstanley, P. Healy: Comparing Learning Strategies for Topographic Object Classification, *IEEE International Geoscience and Remote Sensing Symposium* (IGARSS'03), July 2003.
[14] L. Keyes and Winstanley AC. Automatically Structuring Archaeological Features on Topographic Maps. *GIS Research UK*, 191-4, Sheffield, April 2002.

[Please continue on next page and further page if necessary]

Project Schedule / Timescale (indicate likely steps in project and timing in **graphical format**)

	Year 1			Year 2		
Sept-Oct	Oct-Dec	Jan-March	March-Sept	Oct-Dec	Jan-April	May-Aug
Work unit 1						
	Work Unit 2					
		Work Unit 3				
		Work Unit 4				
		Academic papers				
			Work Unit 5			
			Presentation	Work Unit 6		
					Academic Papers	
						Dissertation
						Presentation

Project Summarised Work Plan*

Work unit 1 Research & literature review
Tasks -searching for & retrieving research material
-evaluate research documentation

Work unit 2 Compilation of corpora of data, training and test data sets
Tasks -scanning documents
-vectorisation (scan2CAD)
-data clean-up
-object extraction and segmentation (closed polygonal data objects) (autoCAD)

Work unit 3 shape/structural descriptors and recognition
Tasks -Re-engineering existing software (Matlab)
-perform unsupervised learning & clustering on system
-perform supervised learning on system
-database constructs
-Fuse results

Work unit 4 SLM recognition
Tasks -SLM theory
-development of N-gram models
- develop SGLM models
-development of evaluation metrics
-evaluation tool design and construction
-evaluate SGLM models

Work unit 5 Labelling components
Tasks -database construction

Work unit 6 Software system
Tasks -sofware modules design and development

*Group meetings and liaison between ITB, NUIM and Entropic continue throughout each year.

Plans to Disseminate Outcomes

Outputs from this work will include a

- Masters dissertation
- Journal publication
- ITB journal publication
- Conferences and publications:
    - ITB conference, 2005 and 2006
    - 8th International Conference on Document Analysis and Recognition (ICDAR 2005, Seoul) and
    - 6th International Workshop on Graphics Recognition (GREC 2005, Hong Kong), July 2005.
- Graphics recognition software system