Corpus-Based Thesaurus Construction for Image Retrieval in Specialist Domains

Khurshid Ahmad, Mariam Tariq, Bogdan Vrusias and Chris Handy

Department of Computing, School of Electronics and Physical Sciences, University of Surrey, Guildford, GU2 7XH, United Kingdom
{k.ahmad, m.tariq, b.vrusias, c.j.handy}@surrey.ac.uk

Abstract. This paper explores the use of texts that are related to an image collection, also known as *collateral texts*, for building thesauri in specialist domains to aid in image retrieval. Corpus linguistic and information extraction methods are used for identifying key terms and conceptual relationships in specialist texts that may be used for query expansion purposes. The specialist domain context imposes certain constraints on the language used in the texts, which makes the texts computationally more tractable. The effectiveness of such an approach is demonstrated through a prototype system that has been developed for the storage and retrieval of images and texts, applied in the forensic science domain.

1 Introduction

A visual scene in a specialist domain may contain a range of information that usually cannot be detected by an untrained person. An art critic can discern several aspects of an object in a Cubist painting and convince us that what appears to be a juxtaposition of geometrical elements is a striking portrait, or a bicycle, or indeed a group of soldiers creating mayhem. An experienced scene of crime officer can identify and describe the 'existence' of various (parts of) human beings and objects in a scene-of-crime image, particularly the physical attributes and relative locations of these objects, which may not be obvious to 12 adults sitting on a jury panel. However, once the officer describes the objects, the untrained person can generally immediately discern the attributes and locations that, hitherto, were not so obvious to them. The *link* between an image and its verbal description is one of the central issues in most disciplines that deal with human vision and by implication human intelligence. The Scene of Crime Information System (SoCIS) project[1] has attempted to explore this link by analyzing the use of texts related to images, also known as *collateral texts*, for the indexing and retrieval of specialist images. We have adopted a corpus-based approach and used information extraction techniques in developing a prototype text-enhanced image storage and retrieval system. This paper reports work in progress on

[1] A three-year EPSRC sponsored project (Grant No.GR/M89041) jointly undertaken by the Universities of Sheffield and Surrey and supported by five police forces in the UK.

the construction of a thesaurus from domain-specific texts for query expansion purposes.

The development of digital visual archives brings with it the problem of indexing the images in the archives. These indices act as the equivalent of keywords used to index text documents. The currently available digital archives range from medical image archives to archives of the images of paintings and press-agency photo-collections. During the last three decades significant effort has been spent on systems that focus exclusively on vision-specific features such as colour distribution, shape and texture, often called *content based image retrieval systems (CBIR)*, and this term can be used to describe many research as well as commercially available systems [1]. However, CBIR systems have an implicit limitation in that visual properties cannot, in themselves, be used to identify arbitrary classes of objects. Indeed there are theoretical limitations on using the visual features for describing an image [2] and, recently, some practical limitations of such an approach have been outlined as well [3]. Earlier image retrieval systems, as well as the images retrieved by the search engines, rely almost exclusively on keywords. The problem here is that appending keywords to an image is not only quite time consuming, the estimates vary from minutes to hours [4][5], but the choice of keywords may in themselves show the bias of the indexer. This has led to the so-called *multimodal* systems, which essentially use linguistic features extracted from textual captions or descriptions together with the visual features for storing and retrieving images in databases. The terminology used in the description of such systems is indicative of the multimodal nature: Picard's *Visual Thesaurus* [6]; Srihari's arguments on 'texts' that are *collateral* to an image [7].

Retrieval is shown to be more effective when textual features are used together with the visual features, for example [8] show a mild improvement on precision and recall statistics when the combined features were used to query an image database compared to when either text or visual features were used, but there are still limitations where the keywords are concerned in that the use of synonyms, abbreviations or related words as well as broader or narrower words is not taken into account. The issue of *inter-indexer variability* [4], the variation in the verbal outputs of different indexers for the same image, has shown a use of related terms. The term *query expansion*[2] refers to the addition of search terms to a user's search for improving precision and/or recall. The additional terms may be either taken from a thesaurus or from documents that the user has specified as being relevant. A thesaurus is a "controlled and dynamic vocabulary of semantically and generically related terms, which covers a specific domain of knowledge"[9]. The most common relationships in a thesaurus are related terms (RT), broader terms (BT) and narrower terms (NT). There do exist some general purpose thesauri, Roget's[10] thesaurus being the classic example, as well as lexical resources such as WordNet[3] but the problem is that their coverage is too broad rendering them inadequate for use in specialized domains such as forensic science or medicine. Hence when images of a specialist nature are to be stored there is a problem in that the thesauri and relevant documents for query expansion are not readily available establishing the need to create domain-specific thesauri. These could be manually built by expert

[2] http://wombat.doc.ic.ac.uk/foldoc/ (Site visited 13 Nov 2002)
[3] http://www.cogsci.princeton.edu/~wn/

lexicographers, of which there are a number of examples like the NASA thesaurus[4] and the Arts and Architecture Thesaurus (AAT)[5], however, handcrafted thesauri face similar problems to those of manual keyword indexing of being time-consuming to build, subjective and error-prone, as well as having the additional issue of inadequate domain coverage. A solution to this is the automatic generation of thesauri for specialized domains from representative text documents.

Automatic thesauri generation was initially addressed by [11][12], as far back as the 1970s, through the use of association matrices, which use statistical term to term co-occurrence measures as a basis for identifying related terms. This method has a number of drawbacks: many unrelated terms will co-occur due to being highly frequent or general; synonyms are hardly used together; only single-word terms are considered whereas in a number of specialist domains multi-word terms are used frequently; a cluster of associated terms is produced with no knowledge of the kinds of relationships between the terms. [13] addressed the fact that synonyms were more likely to have similar co-occurrence patterns rather then co-occur in a document or document collection, by associating a term with a phrase based on its contextual information. The SEXTANT system [14] uses weak syntactic analysis methods on texts to generate thesauri under the assumption that similar terms will appear in similar syntactic relationships, and groups them according to the grammatical context in which they appear. Both the methods above are viable approaches but still do not address the shortcoming of undefined relationships between terms.

In the hope of addressing some of the issues mentioned above we intend to explore a different approach to thesauri construction -based on the context of specialist languages within recent developments in corpus linguistics, and in particular corpus-based lexicography and corpus-based terminology. The proponents of corpus linguistics claim that a text corpus, a randomly-selected but systematically organized collection of texts, can be used to derive empirical knowledge about language, which can supplement, and frequently supplant, information from reference sources and introspection [15]. A significant practical application of this empirical approach has been found in dictionary making or lexicography: this is facilitated in large measure by the computation of the frequency and collocation of tokens in the text. Expert lexicographers can then elaborate on the elicited vocabulary [16]. The random selection involves selecting equal chunks from every text collected or selecting randomly from a catalogue of 'books in print.' The systematic organization refers to selecting different genres of texts – formal and informal types, for example journal articles and popular science articles, instructive and informative types, for example, advanced texts and basic textbooks and instruction manuals, and so on. There is much discussion in corpus linguistics about what constitutes a *representative* corpus. This indeed is an issue for texts written in general language or language of everyday use. If one wishes to use the methods of corpus linguistics for specialized subjects, the question of representativeness is not as vexatious. This is, perhaps, because the linguistic output of a specialist community is limited, in sheer volume and in genre, as compared to that of the broader general language community.

[4] http://www.sti.nasa.gov/thesfrm1.htm
[5] http://www.getty.edu/research/tools/vocabulary/aat/

Specialist languages, considered variants of natural language [17], are restricted syntactically and semantically, which makes them easier to process at the lexical, morphological and semantic levels [17, 18, 19]. There is a preponderance of open class words in specialist languages, particularly nouns and nominalizations, as they deal with objects and named events, actions and states. Specialist languages tend to use a large number of compound terms and these compounds also relate to named entities within the domain. Again, as science and technology deals with named entities researchers aim to create structures to organize the interrelationships between these named entities. This organization is argued for and reported in the literature of the domain, through the use of lexical semantic relationships. It has been suggested that not only can one extract keywords from a specialist corpus [18, 20, 21,] but one can also extract semantic relations of taxonomy and meronymy (part-whole relations) from free texts [22]. Thus, for us, the extraction of terms and their interrelationships from a text corpus to start building a thesaurus is an attractive proposition. Section 2 of this paper discusses the issue of the association between images and texts under the idiosyncratic term *collateral texts* and how one goes about building a representative corpus of collateral texts in order to construct a thesaurus. The issue of domain coverage has been investigated through the comparison of terms extracted from a *progeny* corpus (representative of a sub-domain) to those extracted from a *mother* corpus. This functionality has been incorporated in a text-enhanced image retrieval system – (Section 3). A brief outline of on-going work is given in Section 4.

2 Thesauri Construction from Specialist Text Corpora

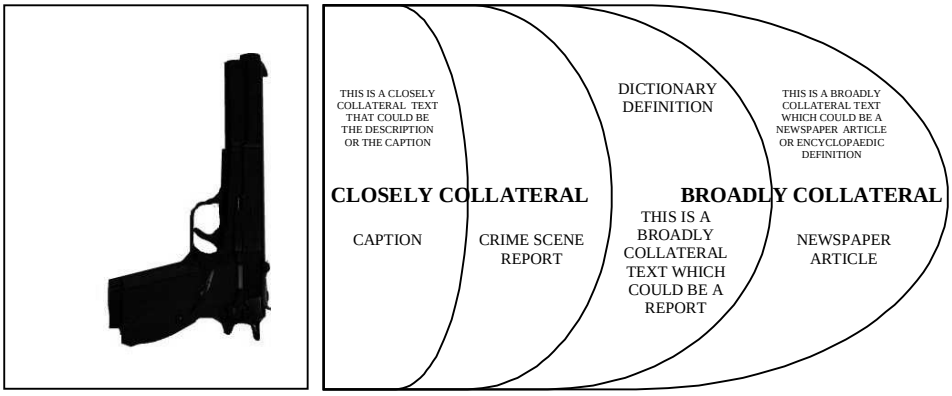

Fig. 1. Closely and broadly collateral texts

An image may be associated in various ways with the different texts that may exist *collateral* to it. These texts may contain a full or partial description of the content of the image or they may contain metadata information. Texts could be *closely* collateral like the caption of an image which will describe only what is depicted in the image, or

broadly collateral such as the report describing a crime scene where the content of the image would be discussed together with other information related to the crime; the closer the collateral text to the image the higher the co-dependency. The degree of co-dependence between an image and its collateral text can be exploited for indexing and retrieving images at different levels of abstraction.

The computer-based analysis of closely collateral texts, written/spoken for a restricted readership, say by scene of crime officers for describing scene of crime images, will help in the identification of objects and their relationships in an image of a specific scene of crime. This information can then be used for indexing that particular image. An analysis of broadly collateral texts, texts written to instruct and inform a broader readership, for instance a scene of crime manual dealing with the collection of evidence written by experts for scene-of-crime officers or a new technique developed in a journal paper, will help in the identification and elaboration of broader terms of the domain, which can be useful in the construction of a thesaurus. Such a thesaurus has to be updated at regular intervals or indeed very frequently in specialisms that are undergoing rapid change. Forensic science may be a good example – here developments in computer vision, in analytical chemistry and molecular biology together with developments in law regarding indirectly-derived evidence, for example, DNA fingerprinting, and digital photography, not only affect the administration of justice but bring a plethora of new terms that have been adopted by forensic scientists and used by police officers. Many of these terms are introduced in the broadly collateral texts and, in due course, find their way into closely collateral texts.

A corpus-based approach can be used to investigate the language of forensic science for thesaurus building purposes, where the corpus can be said to consist of broadly and closely collateral texts with respect to a typical scene of crime image collection. The aim is to study the behavior of the language at the lexical, morphological and lexical-syntactic levels to determine whether it has a discernable structure that can be used to extract terms and their relationships. Typically, following work on corpus-based lexicography [15], the terminologists collect and analyze a (random) sample of free texts in a given domain. The text is analyzed at the lexical and morphological level and the frequency of tokens and their morphological forms is noted. Candidate terms are produced by contrasting the frequency of tokens in the specialist corpus with that of the frequency of the same tokens in a representative corpus of general language. This ratio, sometimes referred to as a *weirdness* measure [18], is a good indication that the candidate will be approved by an expert to be a term.

$$\text{weirdness coefficient} = \frac{f_s / N_s}{f_g / N_g} \quad \textbf{(1)}$$

Where f_s = frequency of term in a specialist corpus
f_g = frequency of term in general language
N_s = total number of terms in the specialist corpus
And N_g = total number of terms in the general language

A forensic science corpus of over half a million words has been created. To ensure that the corpus is representative of the domain, a variety of text types ranging from

1990-2001 were used. Our corpus comprises 1451 texts (891 written in British English and 560 in American English) containing a total of 610,197 tokens. The genres include informative texts, like journal papers, instructive texts, for example, handbooks and imaginative texts, primarily advertisements. These texts were gathered from the Web using the Google Search Engine by keying in the terms *forensic* and *science*. We analyzed the frequency distribution of compound terms in the mother corpus and consulted our expert informants –scene of crime officers (SOCOs) regarding more specialized texts: We found two sub-domains, *crime scene photography* (CSP) and *footwear impressions* (FI) and collected texts from the Web by keying in the two terms and following the links indicated by the texts. (The CSP corpus has 63328 tokens and the FI 11332). Furthermore, the SOCOs provided 53 crime-scene forms comprising 6580 tokens. As part of the SoCIS project, 10 SOCOs provided a subsequently transcribed commentary on 66 scene-of-crime images comprising a total of 5,000 tokens. In this way we had built a mother corpus of Forensic Science and four progeny corpora representative of sub-domains of Forensic Science that could be used for a comparative analysis.

2.1 Lexical Signature of the Domain

The major building blocks of a thesaurus are the individual words, lexical units or terms that form its backbone. The frequency of occurrence of open class words (OCWs) within a corpus can be an indication of terms that are accepted as part of that language's register. A frequency analysis was conducted on the forensic science corpus to determine its *lexical signature*: the key single words used frequently to situate the text in the domain of forensic science. Typically, the first hundred most frequent tokens in a text corpus comprise over 40% of the total text: this is true of the 100-million word British National Corpus (BNC) [23], the Longman Corpus of Contemporary English (LCCE), as is true of a number of specialist corpora we have built over the last 15 years [20]. The key difference is that for the general language corpora the first 100 most frequent tokens are essentially the so-called closed class words (CCW) or grammatical words: in the specialist corpora, in contrast, as much as 20% of the first hundred most frequent tokens comprise the so-called open-class or lexical words. The import of this finding, for us, is that these frequent words are used more productively in the morphological sense in that their inflections (plurals mainly) and compounds based on these frequent words also tend to dominate the text corpus. The frequent use attests to the acceptability of the single and compound words: this for us is crucial in building a thesaurus. A look at the 100 most frequent words in the Forensic Science corpus shows that the first 20 most frequent words are indeed the CCWs comprising just under 30% of the total corpus – a figure similar to the BNC. The next 10 most frequent tokens comprise three open class words –*evidence*, *crime*, and *scene*: these 10 words comprise 3.78% of the total corpus and three open class words contribute a 1.0% to this 3.78%. In itself 1% is a small number, but studies of word frequency suggest otherwise: for every set of 100 words in a forensic text it is statistically possible that 1 word would be either of the three. The total contribution of the 21 open class words amongst the 100 most frequent comes to about 6% - this is the frequency of the most frequent word in written English – the determiner *the*.

The arguments of Halliday and Martin[17] can also be partially attested by noting four derivations in Table 2 (*identification*, *analysis*, *information*, and *investigation* from the verbs *to identify*, *to analyse*, *to inform* and *to investigate*). The lexical signature can be identified more vividly by comparing the distribution of these tokens in the BNC. The CCWs are used with the same (relative) frequency, but the open class occur far more frequently – the token *forensic* is 471 time more frequent in our corpus as compared to the BNC, followed by *crime* (53 times), *scene* (38 times), and evidence (20 times).

Table 1. Lexical Signature of the Forensic Science Corpus

Tokens	Cumulative Relative Frequency
the,of,and,to,a,in,is,be,that,for	23.30%
or,on,as,was,by,s,with,are,from,it	5.97%
this,an,*evidence*(23),at,not,*crime*(26),can,have,which,*scene*(30)	3.78%
were,*forensic*(32),should,he,will,when,*police*(37),may,if,de	2.39%
has,been,other,one,they,all,*identification*(47),had,used,these	1.97%
but,*case*(52),also,their,his,there,any,*found*(58),*court*(59),such	1.59%
two,*analysis*(62),more,what,*body*(65),i,no,who,use,d	1.35%
some,where,*blood*(73),*time*(74),*information*(75),l,you,only,into,*victim*(80)	1.19%
must,m,*dna*(83),would,her,then,*science*(87),*sample*(88),most,than	1.06%
we,*cases*(92),after,*test*(94),made,about,*investigation*(96),its,new,each	0.98%

Weirdness shows the skewness in the distribution of the words in two corpora. A higher weirdness indicates significant use of the word in the specialist corpora as compared to general language corpora, an extreme example being a weirdness of infinity indicating that the term is not present in the BNC at all. This might be indicative of neologisms in the text, which have not yet been adopted in general language. The frequency ratio is a good indicator of *termhood*: Table 2 lists a number of terms that have *infinite* weirdness; note some are single word technical terms, but many are *compounded* terms like *bitemark*, *toolmark* or plurals of terms like *shoeprints* the singular of which does exist in the BNC.

Table 2. Terms with a weirdness of infinity ordered on relative frequency (*f/N*), N = 610,197

Single Term	*f/N*	Compound Term	*f/N*	Compound Term	*f/N*
rifling	0.0139%	bitemark	0.0174%	spectroscopy	0.0092%
pyrolysis	0.0124%	earprint	0.0122%	handguns	0.0090%
accelerant	0.0105%	nightlead	0.0105%	shoeprints	0.0070%
polygraph	0.0081%	handgun	0.0105%	toolmark	0.0045%
accelerants	0.0079%	fingerprinting	0.0093%	earprints	0.0040%

Tokens that are absent in the BNC are a part of the potential signature of a specialist domain. The other tokens that form part of the signature are tokens with significantly high relative frequency – the italicised open class words in Table 1. These tokens/terms are used productively to make compound terms; something which we show is cruical for a thesaurus to be used in query expansion. Table 3 comprises 10 of the frequently used tokens that are amongst 100 most used open class words together with an exemplar set of (relatively) high frequency compounds that comprise two of the 10 single tokens.

Table 3. Highly frequent terms and their compounds

Mother Corpus			Total 610,197
	f	*f/N*	*Weirdness*
analysis	862	0.0014	10.54
blood	781	0.0013	12.43
crime	2366	0.0038	53.52
dna	676	0.0011	33.05
evidence	2757	0.0045	20.77
forensic	1563	0.0025	471.04
homicide	237	0.0004	228.30
physical	382	0.0006	6.45
scene	1605	0.0026	38.18
science	634	0.0010	9.68

Compound	Singular	Plural
blood spatter	17	12
crime scene	495	69
dna analysis	39	Not found
forensic science	229	82
physical evidence	161	NA

Indeed, the two most frequent tokens, *crime* and *scene* are used to form over 90 different compounds, some comprising upto three other (high frequency) tokens, for instance *crime scene* investigator, *crime scene* photography, *crime scene* processing, *crime scene* technician, *crime scene* sketch, and *crime scene photography personnel agency*. Much the same is true of other tokens in Table 3: *blood*, *dna* and so on are used just as productively. Note also the tendency of the authors of these texts to use plurals of terms as well as the singulars. The identification of a lexical signature for a domain, containing both single and compound terms, from a randomly selected corpus will help to initiate the development of a thesaurus for the domain.

The weirdness measure can also be used to determine the lexical signature of a *progeny* corpus as compared to a *mother* corpus. In the above example the mother corpus was the BNC and the forensic science corpus the progeny corpus. A statistical sampling was done of the OCWs for these three corpora in comparison to the mother corpus. The ratio *r* refers the relative frequency, *f/N*, where *f* is the frequency of the token and *N* the total number of tokens. Note that the highly frequent OCWs in the *Footwear* and *CS Photography* progeny corpora generally have a much higher weirdness when compared to the FS mother corpus then that of the same words in the FS corpus compared to the BNC. Table 4 shows a comparison of selected high weirdness terms in the three progeny corpora. The analysis of progeny corpora may yield more specialized terms, indicating they have their own lexical signature, and these either could be added to the main thesaurus or kept separately as a sub-domain thesaurus.

Table 4. Comparison of three progeny corpora

Footwear	r_F/r_M	r_m/r_{BNC}	CS Photos	r_{CS}/r_M	r_{CS}/r_{BNC}	Transcripts	r_T/r_M	r_T/r_{BNC}
(a)	***(b)***	***(c)***	***(d)***	***(e)***	***(f)***	***(g)***	***(h)***	***(i)***
footwear	126	40	lens	67	8	footwear	41	40
reebok	55	2	underexposed	49	INF	fingermarks	18	INF
molding	55	INF	lenses	43	3	ricochet	18	19
gatekeeping	27	INF	tripod	43	12	strangulation	16	21
impressions	25	31	enlargements	39	17	splatter	16	35

Single word terms, with some key exceptions, are often used as *carrier terms* in that they form compounds to give more specific meaning to an existing concept: in themselves the single word terms are usually too generic. An automatic extraction and

analysis of compound terms may perhaps lead to the conceptual structure of the domain in question. Compound terms typically have a nominal head qualified by an adjective or compounded with another noun or noun phrase. They are usually not interspersed by closed class words. For example, *gunshot residue analysis tool* may be written as *a tool for the analysis of residues left by a gunshot* but that will be rare.

NP ➞ [adjective] | [Noun], NP

A heuristic to identify compound nouns in free text is as follows: given that we have a sequence of two or more consecutive words it is possible to assume that this sequence of words is a compound term provided that there are no closed class words in the sequence. For development of our method for extracting terms for an arbitrary domain this is another important heuristic for searching for compound terms and perhaps also for validating them. Church and Mercer have suggested that one may use the Student's *t-test* for testing whether or not a collocation, or a compound token found in the analysis of a text or corpus, is due to random chance [24]. The authors have simplified the computation of the *t-test* score and suggest that for a collocate *x+y:*

$$t \approx \frac{f(x,y) - f(x) * f(y)/N}{\sqrt{f(x,y)}} \quad (2)$$

where *f(x,y)* is the frequency of the compound token, and *f(x)* and *f(y)* that of the single tokens comprising the compound. Church and Mercer have suggested that 'if the *t-score* is larger than 1.65 standard deviations then we ought to believe that the co-occurrences are significant and we can reject the null hypothesis with 95% confidence though in practice we might look for a t-score of 2 or more standard deviations'. In Table 5 we show the *t*-scores for a number of high frequency compound tokens in the FS corpus. We have tabulated both the singular and plural forms of the token where relevant; hence the ~ sign, for example, for *scene* indicating both singular and plurals.

Table 5. 10 Most frequent compound terms

x	*y*	*f(x,y)/N*	*f(x)/N*	*f(y)/N*	*t(x,y)*	*t(x,y)/2*
crime	scene~	0.000912	0.003825	0.002858	23.46	11.73
forensic	science~	0.000503	0.002527	0.00124	17.53	8.76
workplace	homicide	0.000145	0.00032	0.000462	9.48	4.74
crime	lab(orator)~	0.000149	0.003825	0.00167	9.18	4.59
cartridge	case~	0.000131	0.000414	0.002656	8.92	4.46
body	fluid~	0.000112	0.001352	0.000236	8.28	4.14
criminal	justice	9.21E-05	0.000689	0.00047	7.52	3.76
fire	scene~	8.89E-05	0.000774	0.002858	7.23	3.62
dna	analysis	6.3E-05	0.001093	0.001394	6.09	3.05
blood	spatter~	4.69E-05	0.001263	9.7E-05	5.37	2.69

2.2 Discovery of Conceptual Relations

Every language has its own vocabulary, where lexical units or words represent certain concepts and these words are grammatically arranged in certain patterns that convey a

meaning. A range of semantic relations may exist between these different lexical units. [25] presents a model that illustrates some basic relationships between classes of entities: *Identity*, where class X and class Y have exactly the same members. The lexical relation which corresponds to this is *synonymy*, for example "fingerprint" and "lift" are synonyms in that they are syntactically equal; *Inclusion*, where class Y is entirely included within class X. The lexical relation corresponding to this is *hyponymy*, which is most commonly illustrated by the construct 'Y is a kind/type of X.' An example of this could be 'a gun is a type of firearm'. The most common type of lexical hierarchy is a *taxonomy*, which reflects the hyponymy relationship also known as the supertype/subtype or subsumption relationship and a *meronomy*, which models the part-whole relationship. This section attempts to discover frequently occurring patterns that demonstrate these two types of relationships.

Phrasal structures such as compound words convey a certain semantic relationship between the constituent lexical units. Usually headwords such as *scene* are weak semantically in that their meaning cannot be easily ascertained out of context unless they are part of a compound, for example crime scene or movie scene specify what type of scene it is. Compounding tends to specialize the meaning of the headword, which can be used to create a hierarchy. For example taking the three compounds blood sample, fingerprint sample and DNA sample, it shows that *blood*, *dna* and *fingerprint* are different types of *samples*. Similarly there may be certain lexical cues such as *kind of* and *part of*, which convey hyponymic or meronymic relationships between certain lexical units they are syntactically associated with. In the following we discuss the possible elicitation of conceptual structures from texts, which can be used to define broader and narrower terms in the thesaurus.

The hypernymy/hyponymy relationship, also known as the supertype/subtype or subsumption relationship, is the semantic relationship that is used to build taxonomies for various purposes –a classic example being the biological classification of species. At the higher levels of a taxonomy more general/broader concepts are encountered, for example knife is a general concept for a dagger or stiletto. There are a number of linguistic patterns that can be used to illustrate the hyponymic relationship in texts. The cues *is a*, or *is a type of* are the most common patterns but there are a number of others that are typically associated with this relationship.

There are certain enumerative cues [22] that can be used to derive hyponymic relationships as well. For example taking the sentence "All automatic weapons *such as* machine guns must be registered" one can derive hyponym ('machine gun', 'automatic weapon'). Typical hyponymic and enumerative cues listed in the literature on lexical semantic analysis include:

Table 6. List of hyponymic and enumerative cues

Hyponymic Cues	*is a*; *kind of*; *type of*; *set of*; *class of*; *belongs to*
Enumerative Cues	*like*; *such as*; *such * as*; *or/and other*; *including*; *especially.*

The aim was to study the patterns in which these cues occur as well as find out the proportion of valid phrases returned (i.e. those that depict a hypernym/hyponym relationship or a meronymy/homonymy relationship). The cue *or other* was the most productive with 80% of the elicited phrases being valid. The cue *belongs to* picked up

a single correct sentence ("chrysotile *belongs to* the serpentine group of minerals that are layer silicates") out of only 2 sentences returned. It should be noted that the percentage of valid phrases calculated from the total phrases returned was based on the judgment of this author. It was interesting to note that for the forensic science domain the enumerative cues had around 60% productivity for a total of 1224 clauses comprising the enumerative cue and the potential compound term; while the typical hyponymic cues had only 10% for a total of 400 clauses. A few example sentences extracted from the corpus for some of the cues listed above are:

Table 7. Example of sentences containing lexical-syntactic cues

Trace evidence,	*such as*	hair and fibers, is collected off the body […]
In the case of shootings	*or other*	fatal assaults the forensic pathologist, […] important trace evidence.
[…] to search for latent fingerprints, hairs, fibers, blood	*and other*	bodily fluids.
[…] the investigation of computer crimes	*including*	computer intrusions, component theft and information theft.

Each set of sentences show a certain similar grammatical pattern for example considering the sentences containing the cue *such as*, *such as* is acting as a conjunction between two phrases P1 and P2 such that P1 on the left hand side is the superordinate and P2 on the right hand side is a list of subordinate types. Typically these sentences display a local grammar [26] comprising a NP followed by an adjective *such*, a preposition *as* and a comma separated list of NPs with a coordinating conjunction *and* or *or* appearing before the final NP. The example pattern shown below can be used to validate sentences that contain the *such as* cue, which are representative of the hyponymic relationship.

[P1] [R] [P2]

P1 ➜ NP

P2 ➜{ [NP_1] [,], [NP_2] [,],……..,[NP_{n-1}], [and | or] [NP_n] }

[R] ➜ [such as]

NP ➜ [Adjective] | [Noun], NP

If this pattern is matched then each $NP_i \in$ P2 is a hyponym/NT for P1

Figure 3 below shows the process of tagging and parsing an example sentence containing the cue *such as*. Compound structure analysis, discussed above, was used to elicit that *evidence* is the broader term for *trace evidence*. After the sentence is tagged, regular expressions are used to check that it is a valid pattern. Then the sentence is parsed to extract the hypernym {hyponym list} pairs. These partial structures are then represented in XML to be used by the SoCIS system for query expansion purposes (see section 3). This module has been developed in JAVA and makes use of the MXPOST[6] tagger. The whole process is fully automated.

[6] http://www.cis.upenn.edu/~adwait/statnlp.html

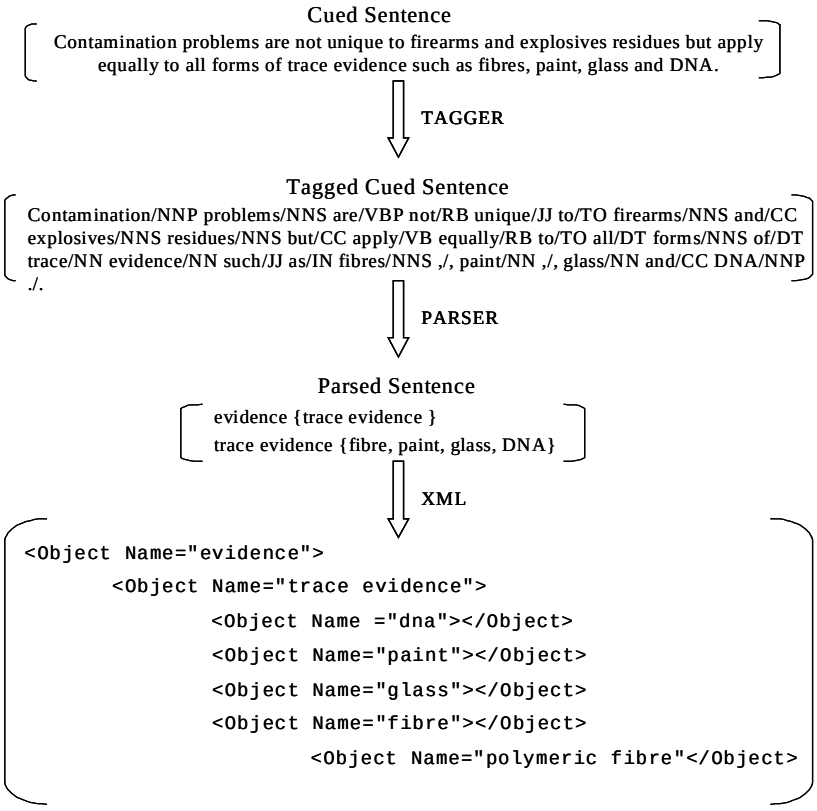

Fig. 3. Analysis of cued sentences

The meronymy/holonymy relationship is most obviously represented by the *has a* cue. Meronymic cues had a productivity of 40% out of a total of 60 retrieved in our corpus indicating that it might not be a commonly used relationship in the domain. The analysis of meronymic cues is treated in the same way as that of the hyponymic cues outlined above.

3 A Text-Enhanced Image Retrieval System -SoCIS

The SoCIS system can be regarded as a variant of image retrieval systems in that it uses not only the vision-specific features of an image but also linguistic features found in a collateral description of the image. SoCIS has three other distinguishing features: First, the system can extract specialist terms and organize the terms in a conceptual hierarchy as discussed in this paper; second, the system attempts to identify meaning-bearing relations, for example “gun on table,” amongst the objects depicted in the image from the verbal descriptions [27]; and, third, the system has a component that can autonomously learn vision-specific features of an image and learn

to associate the keywords used in the description of the image with the vision-specific features [29]. These keywords are extracted automatically from the description using a combination of TF*IDF and weirdness [17] measures as well as filtering carried out using the domain-specific thesaurus. Automation is also a key aspect in SoCIS as most of its features are fully automated. Text analysis and terminology extraction in SOCIS has been facilitated by the integration of two existing systems: System Quirk [17] and GATE [28], which gives SoCIS a powerful text processing mechanism through the combination of a term based statistical approach with a semantic based approach. This system was evaluated using 66 scene of crime images used in the training of scene of crime officers. Experts in training as well as serving officers provided descriptions of the images, and these texts and images formed the input for the system.

A display tool has been developed that can be used to visualize the thesaurus for validating and editing purposes. The XML generated by the method described in the previous section is parsed to display the hierarchies in a tree structure. The user can add, delete or move a node (with its sub-hierarchy unless it is a leaf node) as well as add synonyms. This display tool has been integrated into the SoCIS search interface, a screen shot of which is shown in figure 4. The user can perform interactive query expansion by selecting or deleting nodes as appropriate.

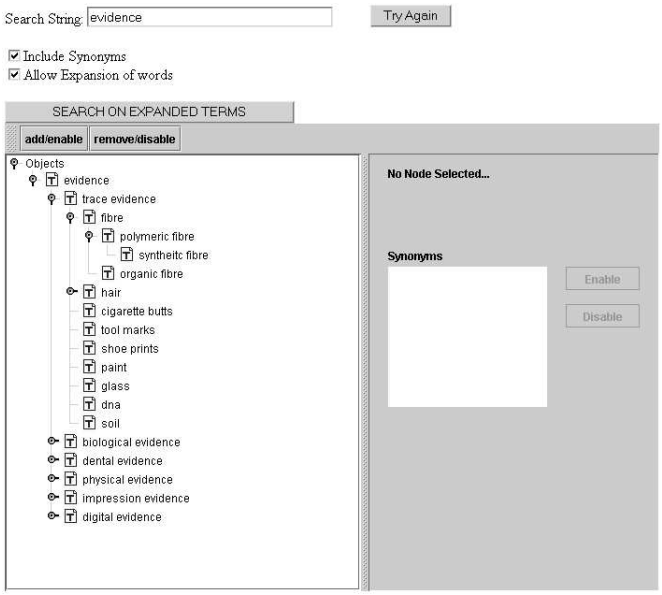

Fig. 4. Screen shot of the search interface

4 Conclusions and Future Work

Content-based image retrieval systems increasingly use keywords collateral to an image for indexing and retrieval. One outstanding problem is that of building thesauri

that can be used during the indexation phase and latterly in the query expansion phase. We have attempted to outline a corpus-based method for building a thesaurus. We have demonstrated, through the use of frequency metrices for single tokens and for compound tokens, that a text corpus, randomly selected and systematically organized, can perhaps be used to initiate the development of such a thesaurus. We have developed a series of programs (in the JAVA programming language) to compute frequency distribution of single and compound tokens and to automatically index an image. The lexical-syntactic pattern analysis carried out has shown that broader and narrow relations can be explicitly extracted which makes it easier to construct a hierarchy and improve the query expansion in perhaps limiting the expansion to narrower terms. From our comparative analysis of progeny and mother corpora it has been shown that inter-variation of terms between sub-domains can be used to study differences in language usage and perhaps one method of ensuring proper domain coverage is for the construction of a domain-specific thesaurus by combining the sub-domain-specific thesauri built from various progeny corpora.

The current method used is based on certain lexical syntactic patterns depictive of hyponymic and meronymic relationships. This work can be continued to discover synonyms as well as other domain-specific relationships such as attempting to identify roles, attributes of entities, and events. In a multi-modal domain it would be interesting to consider a visual representation as well as linguistic labels to represent a concept in the thesaurus, which could act as a link between the two. There has been work done on creating multimedia thesauri but it would be interesting to investigate how images could be automatically analyzed to discover relationships between them based on Picards work [6] and then the visual representation of these images automatically linked to the concept with the corresponding linguistic label. This could perhaps provide even more effective multi-model query expansion.

References

1. Veltkamp, R.C., Tanase M.: Content-Based Image Retrieval Systems: A Survey. (Technical Report UU-CS-2000-34). Institute of Information and Computing Sciences, Univ. of Utrecht, The Netherlands (2000)
2. Marr, D.: Vision. W.H. Freeman, San Francisco (1982)
3. Squire, McG.D., Muller, W., Muller, H., Pun, T.: Content-Based Query of Image databases: Inspirations from Text Retrieval. Pattern Recognition Letters 21. Elsevier Science B.V. (2000) 1193-1198
4. Eakins, J.P., Graham, M.E.: Content-based Image Retrieval: A Report to the JISC Technology Applications Programme. Image Data Research Institute Newcastle, Northumbria. (1999). (http://www.unn.ac.uk/iidr/report.html, visited 19/06/02)
5. Ogle, V.E., Stonebraker, M.: Chabot: retrieval from a relational database of images. IEEE Computer Magazine, Vol. 28(9). IEEE (1995) 40-48
6. Picard, R.W.: Towards a Visual Thesaurus. In: Ian Ruthven (ed): Springer Verlag Workshops in Computing, MIRO 95, Glasgow, Scotland (1995)
7. Srihari, R.K.: Use of Collateral Text in Understanding Photos. Artificial Intelligence Review, Special Issue on Integrating Language and Vision, Vol. 8 (1995) 409-430
8. Paek, S., Sable C.L., Hatzivassiloglou, V., Jaimes, A., Schiffman, B.H., Chang, S.F., McKeown, K.R.: Integration of Visual and Text-Based Approaches for the Content

Labeling and Classification of Photographs. ACM SIGIR'99 Workshop on Multimedia Indexing and Retrieval, Berkeley, CA (1999)
9. Foskett, D.J.: Thesaurus. In: Sparck Jones, K., Willet, P. (eds.): Readings in Information Retrieval. Morgan Kaufmann Publishers, San Francisco, California (1997) 111-134
10. Roget, P.: Thesaurus of English Words and Phrases. Longmans, Green and Company, London (1911)
11. Salton, G.: Experiments in Automatic Thesauri Construction for Information Retrieval. In Proceedings of the IFIP Congress, Vol. TA-2. Ljubljana, Yoguslavia (1971) 43-49
12. Sparck Jones, K.: Automatic Keyword Classification for Information Retrieval. Butterworths, London, UK (1971)
13. Jing, Y., Croft, W.B.: An Association Thesaurus for Information Retrieval. In: Bretano, F., Seitz, F.: (eds.): Proceedings of the RIAO'94 Conference. CIS-CASSIS, Paris, France (1994) 146-160
14. Grefenstette, G.: Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers, Boston, USA (1994)
15. Leech, G.: The State of the Art in Corpus Linguistics. In: Aijmer, K., Altenberg, B. (eds.): English Corpus Linguistics: In honour of Jan Svartvik. Longman, London (1991)
16. Sinclair, J.McH (ed.): Looking Up. Collins, London, UK (1987) 1-40
17. Halliday, M.A.K., Martin, J.R.: Writing Science: Literacy and Discursive Power. The Falmer Press, London and Washington D.C. (1993)
18. Ahmad, K.: Pragmatics of Specialist Terms and Terminology Management. In: Steffens, P. (ed.): Machine Translation and the Lexicon. Springer-Verlag, Heidelberg (1995) 51-76
19. Harris, Z.S.: Language and Information. In: Nevin, B. (ed.): Computational Linguistics Vol. 14, No.4. Columbia University Press, New York (1988) 87-90
20. Ahmad, K., Rogers, M.A.: Corpus-based terminology extraction. In: Budin, G., Wright S.A. (eds.): Handbook of Terminology Management, Vol.2. John Benjamins Publishers, Amsterdam (2000) 725-760.
21. Bourigault, D., Jacquemin, C., L'Homme, M-C. (eds.): Recent Advances in Computational Terminology. John Benjamins Publishers, Amsterdam (2001)
22. Hearst, M.: Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of the Fourteenth International Conference on Computational Linguistics. Nantes, France. (1992)
23. Leech, G., Rayson, P., Wilson, A.: Word Frequencies in Written and Spoken English: based on the British National Corpus. Pearson Education Limited, Great Britain (2001)
24. Church, K.W., Mercer, R.L: Introduction. In: Armstrong, S. (ed.): Using Large Corpora. The MIT Press, Mass., USA. (1993) 1-24
25. Cruse, D. A.: Lexical Semantics. Cambridge University Press, Avon, Great Britain (1986)
26. Gross, M.: Local grammars and their representation by finite automata. In: Hoey, M. P. (ed.): Data, Description, Discourse. HarperCollins, London (1993) 26-38.
27. Pastra, K., Saggion, H., Wilks, Y.: Extracting Relational Facts for Indexing and Retrieval of Crime-Scene Photographs. To appear in Knowledge-Based Systems (2002)
28. Cunningam, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (2002)
29. Ahmad, K., Vrusias, B., Tariq, M.: Co-operative Neural Networks and 'Integrated' Classification. IJCNN 2002, Honolulu, Hawaii, USA (2002)