Article:

Pastra, K., Saggion, H. and Wilks, Y. (2003) Intelligent indexing of crime scene photographs. IEEE Intelligent Systems, 18 (1). pp. 55-61. ISSN 1541-1672

https://doi.org/10.1109/MIS.2003.1179194

Intelligent Indexing of Crime Scene Photographs

Katerina Pastra, Horacio Saggion, and Yorick Wilks, *University of Sheffield*

The Scene of Crime Information System's automatic image-indexing prototype goes beyond extracting keywords and syntactic relations from captions. The semantic information it gathers gives investigators an intuitive, accurate way to search a database of cases for specific photographic evidence.

Photographs capture time in a unique way; they provide a static representation of a dynamic scene, mirroring its properties at a particular moment. It is precisely this characteristic that renders photographs an essential type of documentation in a domain of considerable social importance: crime investigation. The state in which investigators first find a crime scene, the objects and subjects as well as their spatial relations and conditions, all are crucial for collecting evidence and for drawing conclusions during crime investigation.

Although investigating a crime is a time-consuming process, the crime scene cannot be preserved for long: life must take again its normal course, objects must be removed, the space must be cleaned and cleared. As these inevitable changes occur, the risk of contaminating the scene and destroying possible evidence grows. Therefore, scene-of-crime officers take a series of photographs as soon as they arrive at a crime scene, and they create a photo album for each case. Each photo album's first page is an index consisting of a caption for each photograph or set of photographs numbered in sequence. This visual documentation, the official reports of the scene-of-crime officer's actions, and the evidence collected from the scene are the crime investigator's main sources of information.

However, to retrieve information from past cases or to uncover possible similarities and patterns among cases, current practices rely largely on either the investigator's memory or his or her availability to go through piles of case files and photo albums. During the last decades, law enforcement agencies have made many attempts to bring information technology to bear on crime investigation. In Britain, the British Police Information Technology Organisation (PITO) and various software companies have developed management systems to facilitate the administrative aspects of crime investigation.[1] These systems, currently under pilot testing in various police forces, allow monitoring and control of document flow throughout the investigation, visualization of the sequence of events, automatic population of official reports with verbal information provided by the officers, evidence tracking along the whole custody chain, and task and exhibit management. These systems can also store photographs and other case-related information in a central database and allow their retrieval through case-related keywords. Users trace photographs either through their unique case number or through information specific to the case, such as the scene-of-crime officer's name, the type of offense, the date, and the crime scene location. Indexing and retrieving photographs and other case documentation this way will clearly change current practices and facilitate crime investigation.

However, intelligent support for this task could take investigation itself—rather than its administration—a step further. Intelligent, automatic indexing and retrieval of crime scene photographs is one of the main functions of SOCIS, our research prototype developed within the Scene of Crime Information System project.

The SOCIS scenario

SOCIS is a three-year project funded by the Engineering and Physical Sciences Research Council and undertaken by the University of Sheffield and the University of Surrey in collaboration with an advisory board of four UK police forces—the South Yorkshire Police, Surrey Police, Kent County Constabulary, and Hampshire Constabulary.

The prototype, now in its final development and evaluation phase, applies advanced natural language

1094-7167/03/$17.00 © 2003 IEEE
Published by the IEEE Computer Society

processing techniques to text-based image indexing and retrieval to tackle crime investigation needs effectively and efficiently. In the SOCIS research scenario, scene-of-crime officers of the near future will use digital cameras at the crime scene and store the photographs in a central database along with descriptions (captions) that are either spoken (recorded via a hands-free microphone) or written (typed).

SOCIS takes the captions as input, processes them, and extracts relational facts of the form class1:argument1 RELATION class2:argument2. The triples record both the actual strings denoting the relation's arguments and the classes (superconcepts) to which they belong. SOCIS then uses these triples to index the corresponding photographs. Similarly, to retrieve appropriate photographs, SOCIS processes a user query by extracting triples from it, matching the query triples with the indexing triples, and then presenting the matching photographs to the user. Matching takes place first at the relation level; then SOCIS compares triples expressing the same relation at the argument level. If exact argument matching fails, SOCIS tries to find matches to the semantic expansion of the arguments through the class information. If it still obtains no result, it performs simple argument matching (regardless of the presence of any relation), with semantic broadening where applicable.

The automatic extraction of binary relational templates from captions (and queries) is a novel approach to image indexing and retrieval that arose from the idiosyncrasies of our application domain. Crime scene photograph captions chiefly clarify the relations (spatial or other) between the objects depicted. Captions express these relations through prepositions, space- and relation-denoting verb forms, and other adjuncts—for example, "Blood on road surface" and "View of plastic bag containing plant leaves." Crime scene imagery involves static scenes (rather than events), where each object is defined by its position and relation to another object. Extracting the relation with its arguments and rendering this triple a core indexing construct is necessary to overcome the limitations of existing text-based image indexing and retrieval approaches, which are based on keywords or syntactically oriented logical representations.

Existing approaches fall short

For some existing applications, human annotators assign keywords for image classification manually, following in-house classification schemes. Text-based image retrieval in such cases requires users to become familiar with specific wording for queries—using the "right" key terms to bring back the "right" images. Wherever captions are available, researchers have considered them sources of appropriate keywords, which they have tried to extract automatically using statistical methods. Image indexing based on keyword extraction from captions is the prevailing method in text-based image retrieval systems. In such cases, image retrieval relies on keyword matching and semantic expansion. This approach has achieved some high precision scores but very low recall.[2] However, comparative studies of a wide range of variations of the keyword approach—ranging from pure statistical methods to semantic expansion of the keywords and combinations of these two—indicate that the best scores hardly reach 50-percent precision and recall.[3] Researchers have demonstrated that pure statistical methods (such as term-weighting approaches) can coarsely classify images into general categories—such as indoor scenes and outdoor scenes—with great precision.[4] However, for applications that need accurate indexing based on fine semantic distinctions (and thus, accurate representations of what the images depict), this approach would not be adequate.

The automatic extraction of binary relational templates from captions (and queries) is a novel approach to image indexing and retrieval that arose from the idiosyncrasies of our application domain.

An alternative is to extract logical form representations from the captions; these representations take the form of case grammar constructs that mainly capture syntactic dependencies (such as logical objects, agents, and so on) coupled with concept classification information.[5] This approach follows findings in extraction-based text classification, which indicate that automatically extracted, domain-dependent linguistic expressions and associated semantic features perform very well in text classification.[6] The extraction patterns, consisting of a trigger word, conditions to be met, and case roles, are considered dependent on the syntactic context of tokens. Within this framework, research shows that verb forms and prepositions play key roles in indicating classification term meanings.[7]

However, because image captions are so concise, each word has an extremely high information content; therefore, using keyword-based approaches that ignore both syntactic and semantic information in captions will simply fail to differentiate photographs and will often yield incorrect indexing. Using syntactic relations expressed in captions is definitely more efficient but still often yields incorrect indexing. Consider, for example, the captions "View to the loft" and "View into the loft." These captions describe two different photographs belonging to a single case. The first depicts the exterior of a loft; the other depicts the same loft's interior. If we used the keywords "view" and "loft" to index both these photographs, a search for photographs of loft interiors, for example, would retrieve both photographs. Incorporating the semantic relation between the two keywords in the indexing approach would avoid the confusion. This is exactly what the SOCIS approach does: it distinguishes between view DESTINATION loft and view IN loft. In the first case, the caption indicates that the corresponding photograph depicts the view toward the loft—the "destination" of the camera's eye is the loft, the visual focus of the photograph. In the second case, the caption indicates that the photograph shows what's inside the loft. In making this distinction, logical form representations that use syntactic relations are of no help.

Let's look now at another caption: "Position of baby with bedding removed." Indexing this photograph with the keywords "baby" and "bedding" is obviously a mistake, because the absence of bedding in the photograph is precisely what the caption expresses. A method that extracts logical form representations denoting syntactic relations would handle this problem successfully because it would view "bedding" as the logical object of "removed" and thus strongly couple those two words. Indexing the photograph with such a representation (along with others resulting from the caption's analysis) would allow its retrieval whenever the query submitted evoked a similar representation.

However, this approach lacks coverage,

because it remains strongly bound with the caption's surface linguistic realization. Imagine, for example, that another photograph has the caption, "Position of baby with no bedding." This caption's meaning is not different, at least as far as what the photograph depicts, but the syntactic relations that can be extracted differ substantially. The representation of this second caption could still capture the negation, so the indexing would not be incorrect; however, this approach does not consider the two photographs equivalent. Thus, searching for all photographs that depict, for example, "the deceased with no bedding" would return only the photograph with the caption that expresses the negation through the determiner—not the photograph that uses "removed " in its caption. In addition, if the query expressed negation totally differently, it would yield none of the photographs. In SOCIS, however, all these negation cases would result in a WITHOUT relation: baby WITHOUT bedding.

The SOCIS approach uses syntactic relations and concept classification information and adds to these an extraction layer that tries to capture semantics at a deeper level. This approach differs substantially from other approaches for extracting labeled lexical relations[8] of both paradigmatic (X-HYPERNYM-Y) and syntagmatic (write-MEANS-pen) nature. In fact, SOCIS goes beyond syntagmatic lexical relations to what we could call *pragmatic relations,* which are expressed in a specific text type—image captions.

SOCIS indexing prototype

SOCIS implements a corpus-driven indexing approach that resulted from our thorough study of crime investigation documentation and, in particular, crime scene photograph captions. Collecting and analyzing a caption corpus gave us clues to the extraction patterns that would best serve indexing purposes for crime investigation. The SOCIS indexing prototype is a knowledge-based system. Input to the system is a single caption or set of captions in plain text. Starting from a simple tokenizer, SOCIS goes on to use a sentence splitter, a part-of-speech tagger, a lemmatizer, a named-entity recognition module, a parser, and a discourse interpreter. The discourse interpreter houses the rules for extracting the relational facts, which also use a domain ontology and an associated attribute knowledge base. The prototype outputs a set of indexing terms extracted directly from the captions. The extractor can also infer relations not explicitly mentioned in the text; it includes these inferred triples in the final list of indexing terms. Along with the relation triples, SOCIS also extracts single entities, as do keyword-based approaches. When it can find no relational fact in a caption, it performs keyword extraction alone.

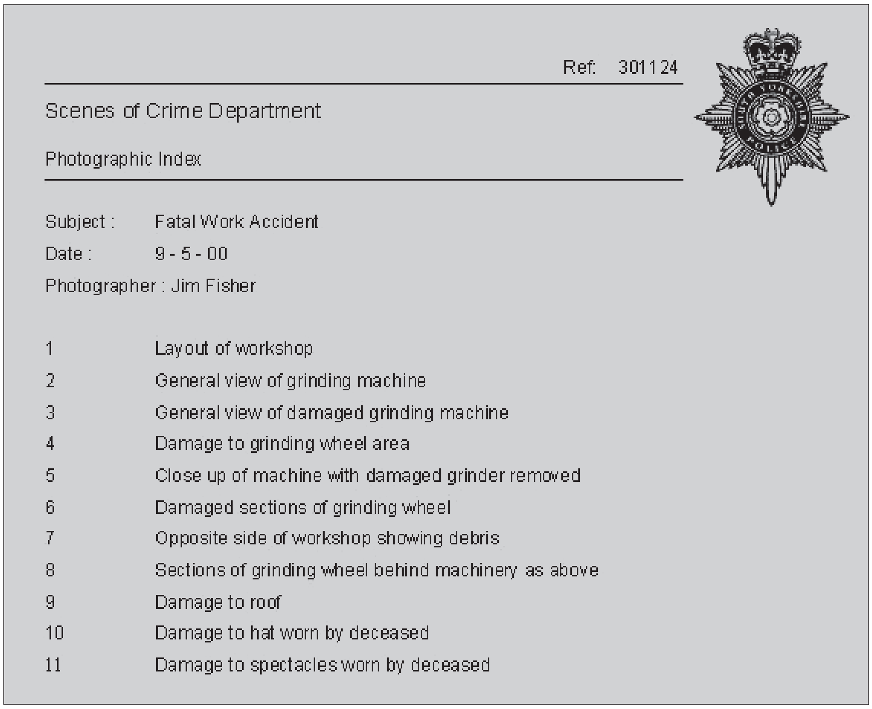
The caption corpus

For the SOCIS project, we collected a corpus of more than 1,200 captions. These captions came from the photo indexes of the albums of more than 350 real crime cases processed at the Rotherham Police Station in South Yorkshire. The vast majority of the captions were written by a single scene-of-crime officer; however, the only significant stylistic difference in the captions produced by different officers is the caption length. We also collected a small set of 65 spoken captions, produced for a SOCIS scenario experiment within a mock crime scene. The experiment, conducted by the University of Surrey research team, involved a Surrey Police scene-of-crime officer attending a murder scene. The officer used a digital camera and recorded a caption for each photograph he took in a digital speech recorder. To avoid (for the moment) automatic speech recognition and transcription problems, we later transcribed these captions manually.

The spoken captions are more verbose than the written ones. Apart from some phenomena inherent in speech (such as *repair*, a repetition to correct misspeaking), the written and spoken captions have many textual characteristics in common: Both kinds of captions are characterized by extensive ellipsis (mainly an absence of verbs) and multiple named entities (such as person and location names), and both kinds mainly refer to what the photographs depict, object properties, and relations. Metainformation is also quite common. Some captions comment on the angle from which the photograph was taken, the photograph's visual focus (such as foreground and background information), or whether it is a distant shot or a close-up.

Figure 1 presents an example photo index from our collection. (To maintain confidentiality, we reproduced this photo index after replacing the dates and person names with fictitious information.)

We initially based the development of our extraction rules on the small set of spoken captions, because these captions are more complex and therefore more demanding and challenging. However, after this first development phase, we expanded and refined the rules by testing them on 500 written captions. Thus, we used half of our corpus for development and the other half for evaluation.



Ref: 301124

Scenes of Crime Department

Photographic Index

Subject : Fatal Work Accident

Date : 9 - 5 - 00

Photographer : Jim Fisher

1	Layout of workshop
2	General view of grinding machine
3	General view of damaged grinding machine
4	Damage to grinding wheel area
5	Close up of machine with damaged grinder removed
6	Damaged sections of grinding wheel
7	Opposite side of workshop showing debris
8	Sections of grinding wheel behind machinery as above
9	Damage to roof
10	Damage to hat worn by deceased
11	Damage to spectacles worn by deceased

Figure 1. Photo index example.

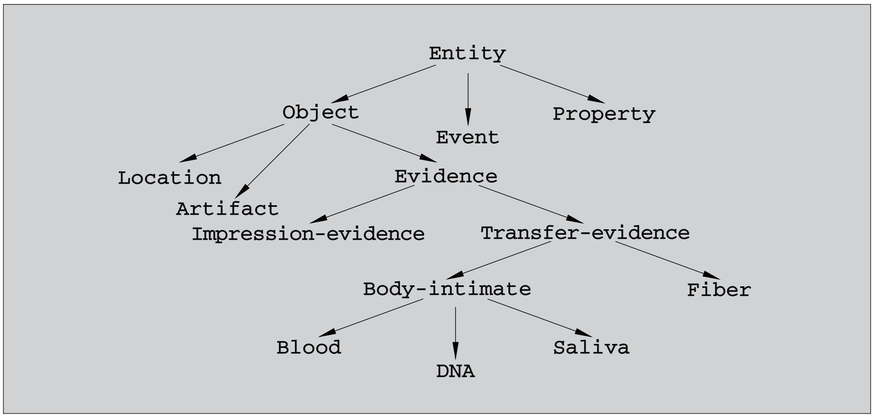


Figure 2. OntoCrime: A graphical representation of part of the ontology.

Text-processing modules

The first four modules on which the SOCIS indexing prototype relies were developed within the GATE project (General Architecture for Text Engineering[9]) and slightly adapted for our application domain. These are the *tokenizer*, *sentence splitter*, *tagger*, and *lemmatizer*.

Once we feed a caption into the system, the tokenizer segments it into words and spaces that provide information on the kind of token (number, word, or punctuation) and its orthographic format. The sentence splitter identifies the boundaries of the sentence that is processed and passes this information to the part-of-speech tagger. Our tagger is a modified Brill tagger[9] that we have tuned for the crime investigation domain. For example, one word we added was "deceased," with the tags NN (noun singular) and VBN (past participle); we put the NN tag first, because our texts use the word almost exclusively as a noun. Some cases required modifications such as the one we made for the word "removed." The default lexicon assigned this token the tags VBD (verb past tense) and VBN (past participle) in this order. So, unless a rule fired that would not allow the VBD tag, the tagger allocated VBD to the token in the text. However, in our captions participles abound and finite verbs are scarce. Therefore, we changed the order of the tags in the tagger's files to give priority to the VBN tag. The last of the four GATE modules is a rule-based lemmatizer, which provides the lemma and the suffixes of each noun and verb found in the caption.

The indexing prototype uses the results of these modules in the named-entity recognition module, which we developed for SOCIS using gazetteer lists and rules expressed in the Java Annotation Pattern Engine notation.[9] We acquired some gazetteer lists from the GATE lexical resources, but we created the vast majority of them from scratch based on lexical information from the PITO lexical database, which was developed for standardization purposes. The module identifies all the types of named entities that might come up in a caption: address, age, conveyance-make, date, drug, gun type, identifier, location, measurement, money, offense, organization, person, time, and other.[1]

The indexing prototype feeds the output of these modules into the next module in the row, the parser. We use an implementation of a common bottom-up chart parser enriched with semantic rules that construct a first-order logical form representation of each caption.[10] This is a robust parser, which means that it allows partial parsing when it cannot construct a syntactic tree spanning the whole sentence. This is important for our application, because most captions consist of phrases that stand as sentences but that don't include an actual verb. The representations that the parser generates consist of a sequence of unary and binary predicates that follow the rules of a context-free phrase grammar of English enriched with features and values. We modified this default grammar to adapt the parser to the nature of the captions.

For example, the parser must handle captions that contain only nonfinite verb phrases and tie together all their complements—for example, "Body on floor surrounded by blood." Originally, the parser generated the following predicates: body(e2), floor(e3), surround(e1), blood(e4), on(e2,e3), by(e1,e4). In the notations identifying each word, e stands for "entity" and the number provides the unique identification for the word. So the predicate on(e2,e3) indicates that e2, which is the token "body," is on e3, the token "floor." The parser also identified two partial syntactic trees, one consisting of the single noun phrase (NP) "body on floor," and another with a verb phrase (VP) consisting of a passive, nonfinite verb phrase (NFVP) and a prepositional phrase (PP), "surrounded by blood." We wanted the parser to come up with a complete parsing of this sentence that would indicate the syntactic relation between the two partial trees. Therefore, we wrote a rule dictating that in every construction of the form NP followed by passive-NFVP followed by PP, the noun phrase that precedes the passive past participle is considered its logical object (lobj), and the whole construction is identified as a verb phrase. With this new rule, the parser also generates the predicate lobj(e1,e2).

Using a model of our domain, the discourse interpreter maps this syntactic representation to a semantic representation.[10] A domain model consists of an ontology (a concept hierarchy) and an attribute knowledge base associated with nodes in the ontology. The discourse interpreter populates the initially bare domain model with instances and relations extracted from the captions during processing, creating a discourse model. When output from the parser is partial, the discourse interpreter tries to complete it according to properties of the identified entities as defined in the attribute knowledge base. We incorporated the SOCIS extractor in this module as an extra processing layer. Using the ontology and the knowledge base, the SOCIS extractor enriches the discourse model with the relational facts of interest and extracts these from the model as indexing terms.

OntoCrime: SOCIS ontology and attribute knowledge base

OntoCrime is the SOCIS domain-dependent ontology, which we developed from scratch using information from the PITO lexical knowledge base and crime investigation practices documentation. We use this domain model mainly to define selection restriction in the triple-extraction phase and to provide class information for the triple arguments. OntoCrime is implemented in the XI knowledge representation language[1] as a direct, acyclic graph with an Entity top node and several Object, Event, and Property classes. These classes have their own subclasses and sub-subclasses going down to the word level. Figure 2 presents a small part of OntoCrime in a tree-like format.

The Object hierarchy consists of a disjunction of classes that denote tangible and intangible objects—for example, Substance, Artifact, Evidence, and so on. The Event hierarchy contains classes denoting state or actions. These include Criminal Actions, such as Assault; Spatial Events, such as Surround; Negation Events, such as Remove; and Metainformation Events, such as Show.

Last, the Property hierarchy has several functional and relational properties that can be assigned to object or event classes through the attribute knowledge base. Simply put, the knowledge base consists of a series of rules that declare the properties of specific classes in OntoCrime. For example, we identified the property can-be-surrounded as a single-valued property in the ontology, and we assigned it to specific object classes through the following rules in the knowledge base:

```
props(conveyance(X),[can_be_surrounded(X,yes)])
props(material_item(X),[can_be_surrounded(X,yes)])
props(impression_evidence(X),
    [can_be_surrounded(X,yes)])
props(fibre(X),[can_be_surrounded(X,yes)])
props(body_part(X),[can_be_surrounded(X,yes)])
props(person(X),[can_be_surrounded(X,yes)])
props(role(X),[can_be_surrounded(X,yes)])
```

The rules dictate that these specific object classes (conveyance, person, and so on) and their children nodes (from subclasses down to the word level, owing to inheritance) are all things that “can be surrounded” by something. Such declarations are very useful for the SOCIS extractor because they serve as semantic constraints for filling in the argument slots of the relational facts.

To illustrate this further, consider the following caption, “Body on floor surrounded by blood.” A rule in the knowledge base declares that the token surround has specific properties:

```
props(surround(X),[(presupposition(X,['Around'
    (AROUND), argument1(AROUND,W),
    argument2(AROUND,T)]):-
   hasprop(X,by(X,W)), hasprop(X,lobj(X,T)),
   (T <– Y ; T <- Y),
   hasprop(Y,can_be_surrounded(_,yes)))]).
```

Whenever the extractor encounters surround in a caption, it presupposes an AROUND relation with two arguments. The first argument should be found in a BY syntactic relation (as produced by the parser). The second argument must be the logical object of surround (also provided by the parser) and be an immediate or indirect instance of a class Y that has the property can-be-surrounded. This way, we make sure that the right arguments accompany the AROUND relation extracted—that is, blood AROUND body and not blood AROUND floor, which is semantically invalid. In this case, our system double-checks the second argument's properties through both syntactic (lobj) and semantic information (the Can-be-surrounded property). In other rules, only one of these kinds of information is available for instructing the system how to choose the right arguments for the triple it extracts.

SOCIS extractor

Following the discourse interpreter's conventions, the SOCIS extractor's rules are written in Prolog. Based on information from OntoCrime and the properties declared in the SOCIS knowledge base, these rules extract 17 different types of relational triples, some of which denote metainformation:

Based on information from OntoCrime and the properties declared in the SOCIS knowledge base, these rules extract 17 different types of relational triples.

- ABOVE. For example, the caption “View of roof above seat …” yields the triple view ABOVE seat.
- AND. This grouping relation functions mainly to imply other relations that hold for all the entities linked with the AND relation. It covers cases of coordination and enumeration. For example, the caption “Bottles, gun, and ashtray on …” yields bottles AND gun, gun AND ashtray.
- AROUND. “Tie around right arm” yields tie AROUND right arm.
- BEHIND. “View of bottles behind the bar” yields bottles BEHIND bar.
- BETWEEN. “Photograph of deceased between vehicle and garage wall” yields deceased BETWEEN vehicle - garage wall.
- DESTINATION. “View of Mansfield Road heading toward Wales Bar” yields Mansfield Road DESTINATION Wales Bar.
- IN. This relation indicates the literal meaning of *in* (inside). For example, “Blood drops inside the bathroom” yields blood drops IN bathroom.
- MADE-OF. “Footwear impression in blood” yields footwear impression MADE-OF blood.
- METAPOSITION. “Shot of bar with tables in the foreground” yields bar WITH tables, tables METAPOSITION foreground.
- NEAR. For example, body NEAR table is denoted via words such as “near” and “adjacent.”
- OF. The extractor uses this relation to express cases denoting a “part-of” relation—for example, rear OF machine.
- ON. “Table showing bottles” yields bottles ON table.
- SOURCE. “Rear garden from Lancing Street” yields rear garden SOURCE Lancing Street.
- SOURCE-BEHIND. This relation denotes the viewpoint from which the photograph was taken. For example, “Shot of floor from behind the bar” yields floor SOURCE-BEHIND bar.
- UNDER. “Chair leg found underneath table” yields chair leg UNDER table.
- WITH. “Bag containing plant leaves” yields bag WITH plant leaves.
- WITHOUT. This relation captures negation or the absence of something. For example, “Table knife with no blood” yields table knife WITHOUT blood.

As these examples show, relation extraction goes beyond the actual presence of prepositions in the captions. The arguments of these relational facts aren't necessarily recorded in OntoCrime; when they aren't classified there, the discourse interpreter adds them under a general class according to their grammatical category; for example, nouns go under the Object class.

The arguments can be any type of object, even named entities, but not metaobjects (such as “shot” or “photograph”), because these have low information content as indexing terms. An argument can be a multiword noun phrase (nouns with various nominal modifiers); in this case, SOCIS extracts triples with both the whole multiword argument and just the head of the noun phrase.

If the relation denoted in the triple is not the OF relation, the extractor also checks the arguments to determine whether they are instances of the Part-denoted class in OntoCrime. It does this to avoid extracting meaningless triples. Consider, for example, the caption, “Fingerprint impression on piece of wood.” In this case, SOCIS extracts the triples fingerprint impression ON wood, impression ON wood, and piece OF wood, but it avoids extracting something like fingerprint impression ON piece.

Furthermore, the extractor filters the arguments so that the two arguments in the triple are never the same. It does this to avoid, among other things, repair cases found in the spoken captions. For example, a spoken caption might say “Shot of hand of left hand with …,” but extracting the triple hand OF left hand would obviously be wrong.

In some cases, instead of describing what the photograph depicts, the caption points to the preceding photograph—for example, “Same shot ….” We have identified some clue anaphora expressions that indicate that triples extracted from the previous caption apply to the current caption as well. SOCIS uses these clue expressions to deal with such cases.

When SOCIS extracts only one relation from each caption, things are quite straightforward. However, most captions consist of more than one relation, and in these cases the system must follow rules to extract the right triples. Our application uses a finite set of rules to assign the appropriate arguments to the triples and to infer other implicit relations. The default rule for cases when many relations exist in a caption is to extract the relations with their arguments in sequence (left-to-right attachment). For example, from the caption “Bottles on table near the bed,” the triples extracted share one argument: bottles ON table, table NEAR bed.

We have captured exceptions to this general rule with more specific rules. For instance, in the caption “Shot of footprint on top of bar,” the default rule would instruct the extractor to extract two triples that share the argument top, which belongs to the Part-denoted class: footprint ON top, top OF bar. But, because the rules allow only the OF relation to have arguments that belong to the Part-denoted class, SOCIS would fail to extract the ON triple. Therefore, we have created a rule that leads to extraction of the ON triple with bar as its second argument: footprint ON bar.

We discovered an even more complicated case in a spoken caption, in which the default rule would also lead to failure: “Photograph from behind bar of body on floor.” In this caption, the default rule would extract the meaningless triple bar OF body. However, our exception rules let the SOCIS extractor avoid such mistakes and extract the right triples: body FROM-BEHIND bar and body ON floor. None of our exception rules is caption-specific; on the contrary, they cover special cases that involve combinations of specific relations and types of arguments.

Besides handling multirelation cases, the SOCIS extractor can also infer triples. We have used the AND relation for inference extensively. First, we defined AND transitivity rules for cases when the caption gives a list of objects (noun phrases)—for example, “Bottles, gun, and ashtray on table.” In this caption, the explicit relational facts are bottles AND gun, gun AND ashtray, and ashtray ON table. However, the AND transitivity rules lead the extractor to infer the bottles AND ashtray triple as well, following mathematical logic.

Furthermore, whenever an entity is shared by two triples, one of which denotes an AND relation, the AND relation’s other argument should also be shared with the other triple. To illustrate, the caption “Broken bottle and door stop found on floor,” contains two explicit relations to be extracted: broken bottle AND door stop, door stop ON floor. However, according to the AND inference rules, the extractor will also extract the implicit triple broken bottle ON floor. This rule does have some exceptions, which we have taken into consideration.

> We have tried to restrict SOCIS’s inference capabilities to cases in which we run a low risk of extracting an incorrect indexing term.

The extractor infers relations in quite a few other cases, one of which we can see in the caption “Photograph of writing in dust on the games machine.” In this case, SOCIS extracts two triples (writing IN dust, dust ON games machine) and infers the triple writing ON games machine. We see a more complex case in the caption “Footprint with zigzag and target on chair.” What the photograph actually depicts is a footprint on a chair; the pattern of the footprint is a zigzag with a target. Apart from the obvious triples that can easily be extracted (footprint WITH zigzag, zigzag AND target, target ON chair), SOCIS infers another three relational facts: footprint WITH target, footprint ON chair, zigzag ON chair. Obviously, the inferred triples carry important complementary indexing information. However, we have tried to restrict SOCIS’s inference capabilities to cases in which we run a low risk of extracting an incorrect indexing term.

Evaluation

The SOCIS indexing prototype is part of a larger system that stores and retrieves crime scene photographs and other case-related documentation, as well as automatically populating official crime scene reports with information that the officers provide verbally. Our additional work on the project includes formal evaluations of the SOCIS system as a whole. We have used half of our corpus of captions for evaluating the SOCIS indexing mechanism alone. We performed a black-box evaluation, in which two colleagues decided whether both the direct and inferred triples extracted automatically by the system were correct or not and whether SOCIS neglected to extract or infer a relational fact. The two judges were in total agreement in their decisions (that is, in characterizing the triples as correct, wrong, or missing); the system scored 80-percent precision and 95-percent recall. This whole process has indicated refinements to our extraction rules and has proved the system’s ability to apply our indexing approach effectively and efficiently.

In addition to this system-oriented evaluation, we have also completed a preliminary user-oriented evaluation of the whole system with real users from the SOCIS advisory board and staff and trainees at the Metropolitan Police Department’s Scientific Support College. We gave these users a developmental version of SOCIS, which had indexed a small database of approximately 100 photographs. Users could retrieve the images by searching using either keywords or relational triples. When the user selected an entity of interest from a drop-down menu, the relations existing in the database that contained this entity as their first argument appeared dynamically in another drop-down menu. When the user selected the relation she wanted, a list of the second arguments of the relation also appeared dynamically. That way, the user could submit a relational fact as a query to the system. Semantic expansion in both the keyword and the relational fact searches could also take place, if the user chose to activate this feature.

The results of this first usability evaluation indicated that indexing images using semantic relations is effective in crime investigation, because these relational facts function as key information in the domain. Searching through relational facts proved not only the most intuitive search method for the users but also the most accurate. A final, more formal overall usability evaluation of the SOCIS system is under way. This time, the system allows free-text queries for image retrieval and returns a weighted list of relevant images, leaving retrieval decisions entirely behind the scenes.

Developing a prototype using real data and users led us to adopt a novel approach to text-based image indexing and retrieval; the users themselves have acknowledged our approach's effectiveness. The SOCIS indexing approach, using advanced natural language processing technology, handles indexing problems that other approaches cannot overcome. However, our approach emerged from work on an idiosyncratic type of text and, in particular, on captions from a specific domain. Would this approach be effective if used for indexing captioned photographs in another domain? If so, how easily could it be ported to this new domain? Considering the knowledge sources we needed to develop our relational-fact extraction mechanism—the ontology, its knowledge base, and the hand-crafted rules—such an effort would certainly encounter the well-known bottlenecks of knowledge-based approaches. On the other hand, porting the approach to another domain would require only minimal domain-dependent modifications to the method's underlying natural language processing technology.

Answering these research questions fully will require the SOCIS indexing method to undergo extensive and thorough testing and experimentation. In the meantime, we hope that our work has made the point that extracting relational facts is not just another alternative image-indexing approach, but one that is indeed necessary in this real-world application.

Acknowledgments

For extensive, fruitful discussions of the project and for their work building the SOCIS system, we thank the Surrey-SOCIS research team members: Khursid Ahmad, Bodgan Vrusias, Mariam Tariq, and Chris Handy. We are grateful also to our police advisory group, and in particular to John Armstrong, head of Scientific Support, Surrey Police; David Ince, principal of the Scientific Support College, Metropolitan Police; and Andrew Hawley, scene-of-crime officer, South Yorkshire Police, for his help both in providing us with information on crime scene documentation practices and in collecting our caption corpus.

References

1. K. Pastra, H. Saggion, and Y. Wilks, *Socis: Scene of Crime Information System,* tech. report CS-01-19, Univ. of Sheffield, UK, 2001.

2. T. Rose et al., "Anvil: A System for the Retrieval of Captioned Images Using NLP Techniques," *Proc. 3rd UK Conf. Image Retrieval* (CIR 2000), Springer-Verlag, Berlin, 2000.

3. A. Smeaton and I. Quigley, "Experiments on Using Semantic Distances between Words in Image Caption Retrieval," *Proc. 19th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval* (SIGIR 96), ACM Press, New York, 1996, pp. 174–180.

4. C. Sable and V. Hatzivassiloglou, "Text-Based Approaches for the Categorization of Images, *Proc. European Conf. Digital Libraries*, Lecture Notes in Computer Science, no. 1696, Springer-Verlag, Berlin, 1999, pp. 19–38.

5. E. Guglielmo and N. Rowe, "Natural Language Retrieval of Images Based on Descriptive Captions," *ACM Trans. Information Systems*, vol. 14, no. 3, 1996, pp. 237–267.

6. E. Riloff and J. Lorenzen, "Extraction-Based Text Categorization: Generating Domain-Specific Role Relationships Automatically," *Natural Language Information Retrieval,* Kluwer Academic Publishers, Dordrecht, Netherlands, 1999, pp. 167–196.

7. E. Riloff, "Little Words Can Make a Big Difference for Text Classification," *Proc. 18th ACM SIGIR Conf.*, ACM Press, New York, 1995, pp. 130–136.

8. S. Richardson, W. Dollan, and L. Vanderwende, "Mindnet: Acquiring and Structuring Semantic Information from Text," *Proc. Assoc. for Computational Linguistics* (ACL), Morgan Kaufmann, San Francisco, 1998, pp. 1098–1102.

9. H. Cunningham et al., "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications," *Proc. 40th Anniversary Meeting Assoc. for Computational Linguistics*, Morgan Kaufmann, San Francisco, 2002, pp. 168–175.

10. K. Humphreys, R. Gaizauskas, and H. Cunningham, *Lasie Technical Specifications,* tech. report, Dept. of Computer Science, Univ. of Sheffield, UK, 2000.

The Authors

Katerina Pastra is a research associate in the Natural Language Processing group and a PhD candidate in the Department of Computer Science at the University of Sheffield. She is also the research coordinator of the Institute for Language, Speech and Hearing (ILASH) there. Her thesis explores the integration of vision and language and in particular the automatic generation of textual descriptions of 3D graphics. Her interests include multimedia indexing and retrieval, multimodal dialogue systems, machine translation, and human–computer interaction. She holds a bachelor's degree in Greek literature and linguistics from the University of Athens and an MSc in machine translation from the University of Manchester Institute of Science and Technology. She is a member of the Association for Computational Linguistics and the International Association for Forensic Linguistics. Contact her at the Dept. of Computer Science, Univ. of Sheffield, 211 Portobello St., S1 4DP, Sheffield, UK; katerina@dcs.shef.ac.uk.

Horacio Saggion is a research associate in the Natural Language Processing group in the Department of Computer Science, University of Sheffield. His main interests in natural language processing are text summarization, shallow natural language processing, text structure, discourse interpretation, and natural language generation. He received his master's degree from the Universidade Estadual de Campinas, Brazil, and his PhD from Université de Montreal, Canada. Contact him at the Dept. of Computer Science, Univ. of Sheffield, 211 Portobello St., S1 4DP, Sheffield, UK; saggion@dcs.shef.ac.uk.

Yorick Wilks is a professor of computer science at the University of Sheffield, head of the Natural Language Processing group, and director of the Institute for Language, Speech and Hearing (ILASH). His research interests include information extraction, dialogue systems, and machine translation. He is a member of the UK's Engineering and Physical Sciences Research Council College of Computing and a fellow of the AAAI and the European Coordinating Committee for Artificial Intelligence. Contact him at the Dept. of Computer Science, Univ. of Sheffield, 211 Portobello St., S1 4DP, Sheffield, UK; yorick@dcs.shef.ac.uk.