Extracting semantic meaning from photographic annotations using a hybrid approach.

Rodrigo Carvalho, Sam Chapman, and Fabio Ciravegna

The University of Sheffield, UK
{rodrigo, sam, fabio}@dcs.shef.ac.uk

Abstract. This paper evaluates singular then hybrid methodologies for extracting semantics relevant to users in cataloguing and searching of personal photographs. It concentrates upon extraction of meaningful concepts within textual annotations focusing around geographical identification, together with references to people and objects concerning each image. A number of approaches are considered; machine learning, rule based and a novel hybrid approach encompassing previous techniques. This evaluation identifies the strengths of the singular approaches and defines rules best suited to differing extractions providing a higher performing hybrid method.

1 Introduction

In recent years digital cameras have become an essential gadget in the household. With the increasing adoption of mobile photography, inexpensive network transmissions, cheap data storage and a decline in physical printing there is an inevitable expanding number of photographs in both public and private digital collections and agrowing need to search over this information. Existing solutions are incomplete as they fail to tackle the needs of users who require retrieval on the conceptual content of individual images which is harder to capture. Automated techniques to extract data from images have been proposed, for example *C*ontent *B*ased *I*mage *R*etrevial techniques, CBIR[9], which index visual artifacts within images. Other techniques focus upon systems to gather user input for the purpose of user directed archival, online photo sharing services are examples. Such systems encourage image reuse and sharing by utilising additional user input, i.e. comments, tags, temporal and categorical groupings and organisation. One issue with such an approach is that users of digital photography often will put minimal effort into this archival process meaning limited potential reuse of the images. This shortfall in available information makes it necessary to make maximal usage of any annotations provided. This paper examines this issue by investigating means to take advantage of minimal photographic descriptions first existing approaches are detailed more fully.

1.1 Existing Approaches

Many approaches aim to address the problem of maximising image reuse. Current techniques focus upon one of three basic approaches, each of which is now detailed briefly.

Image analysis. Image analysis techniques attempt to extract meaning from the pixel content of an image automatically. Veltkamp[9] surveys the state of the art techniques such as face recognition, edge detection, image segmentation, region classification etc. Such techniques however are largely problematic in real world scenarios for two reasons:

1. **Semantic gap** - extracted regions are visual artifacts within pixels and not semantic concepts which users require, for example, *an objects boundary edge* and not semantic entities like *My brothers car*, *dad* or *the eiffel tower*.
2. **Accuracy** - state of the art has an unacceptable precision and recall to be considered useful in that objects and classifications can be frequently misapplied. Barla et al [2] indicate a 20.7% miss-classifcation in rudimentary binary classifications such as cityscape vs non-cityscape.

Improved structured Knowledge Representations. Representing Knowledge in a standard format is of huge importance as it facilitates its reuse. In recent years a number of exchange formats have been developed focusing specifically upon exchanging information regarding digital images. Exif[1] includes detailed camera settings set at the time of digital image capture. Some of this information is of use for retrieval but again suffers from the issue of *semantic gap* where it fails to embed semantic meaning needed by users. Newer standards such as *MPEG-7*[2] provide a mechanism to encode extended information including regional semantic annotations within an image, unfortunately although a format exists for its representation there is as yet no agreed method to obtain the needed annotations. **User (or community) annotation extractions.** Enlisting user support in image classification has had a recent resurgance in popularity following the success of the ESP game[1] and the development of online photo sharing websites such as Flickr, KodakGallery and many more. Such interfaces attempt to empower users to perform individual or collective annotation/archival of digital photographs. One issue with such approaches is that only a small proportion of the population put in reasonable efforts regarding annotation. Given such systems it is imperative to make maximal usage of any photographic annotations as possible. Many attempts have been made to extract maximal meaning from photographic annotations such as[8][7] but most have focussed upon a complete natural language parse which is too costly to scale to a large scale solution.

2 Extraction Focus

The focus of this paper concerns concentrating extraction efforts upon the needs of the user. According to a study performed by Naaman et al[6] the usefulness of various metadata was considered in aiding users to locate their own photos. The cues rated by users as contextually important for recalling images were found in order of importance to be:

[1] *EX*changeable *I*mage file *F*ormat, was created by the Japan Electronic Industries Development Association (JEIDA). Version 2.1 (the first public release) was released June, 1998, and later updated to version 2.2 in April 2002

[2] http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

1) indoors opposed to outdoors pictures,
2) the identity of people within a photo,
3) the location,
4) the event depicted,
5) the number of people,
6) the season,
7) the year,
8) the time of the day,
9) the weather conditions,
10) the date,
11) the mood in which a picture was taken

Further input from industrial sponsors confirmed these features[3]. Some features from the above list could be obtained from sources other than the image annotations themselves. For example image Exif metadata, basic image analysis can determine basic recognition tasks e.g. indoor vs outdoor environments[2] having a a 93% accuracy. Given these issues five key attributes can be proposed, Location, Person, Object, Event, Temporal of which we focus upon extraction of the first three.

- **Location**: a textual location that the image might depict. This includes not only geographical location names but also far less exacting locations such as *home*, *my road*, *my garden* as well as synonyms for place names such as *the big apple*.
- **Person**: people's names or general references to people such as *dad*, *mum*, *brother*.
- **Object**: conceptual objects depicted in an image. This concept was only identified when a term of obvious importance did not fall into any of the previous categories, such as *football* in the description *Dave and his football*.

One way of obtaining such information is via the analysis of textual descriptions about the images within a collection. The following section introduces and discusses a hybrid approach to tackle such issues.

3 A Hybrid Approach

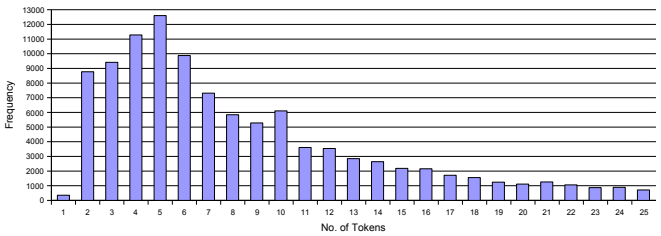
Current research efforts for performing feature extraction from photographs are focused mainly around the solution to widely known computer vision problems. However, with the existence of online photo management and sharing services such as Flickr for almost half a decade, users of this technology have grown accustomed to organising their photo collections by using textual metadata such as single words known as *"tags"* as well as textual descriptions of an arbitrary length. The existence of such metadata about images has opened a window of opportunities for the development of novel techniques for the extraction of information about images by using Natural Language Processing (NLP).

What we propose in this paper is the use of an approach for extracting information from image descriptions that takes advantage of the flexibility of machine learning data models as well as the precision of rule based extractors. Given a very limited initial training dataset as well as a limited number of rules, we aim to combine these two approaches not only for performing more confident extractions from image descriptions, but also to control levels of precision and recall by maintaining a balance over which technique is more influencing in the extractions.

An optimum solution for the domain of image descriptions would have to address two central performance requirements: 1) It must be computationally cheap (*lightweight*) in order to be scalable. 2) The extractions produced must be highly precise

[3] Internal communications with Eastman Kodak Corporation

while maintaining recall. Further to these, performing Information Extraction (IE) from such short snippets of text can be problematic due to their limited grammatical content and disperate presentation. You can see from fig.1 that over 64% of image descriptions have less than 10 tokens in a corpus of over 380K images.



The following sections will introduce the machine learning framework, the rule based extractor and a hybrid approach. The corpus of image descriptions will then be discussed before the evaluation and conclusions.

3.1 Machine Learning

It is widely known that given a small set of training data, machine learning systems are capable of creating a generic model and apply it to previously unseen data. More specifically, in the field of NLP, textual features of tokens (e.g. part of speech, orthography, the tokens themselves, etc.) together with the features of other neighbouring tokens are used in the creation of this model what makes this an extremely flexible technique for extracting information from text.

T-Rex. One such system that achieves competitive results when applied to several corpora is the *T*rainable *R*elation *EX*traction framework (T-Rex)[5]. T-Rex[4] is a highly configurable support vector machine based IE framework that uses a canonical graph-based data model. Its strength comes from decoupling its data representation from the machine learning algorithms allowings configurable extensions.

3.2 Rules

On the opposite end of the spectrum there are rule based extractors that apply manually written Hearst pattern[3] style rules to textual data. Precise extractions can then be performed according to the ganularity of rules.

Saxon. It is a rule based tool for annotating documents and is built upon the Runes 6 framework [4]. It relies on the document being represented as a graph, with nodes representing document elements (tokens, sentences, etc.) and edges representing relationships between elements (belongsTo sentenceXYZ, follows tokenXYZ, etc.). Saxon rules are defined as regular expressions detailing how to move between elements of the graph. A rule has three main parts: a starting point, a regular expression (describing how to move between sections of the graph), an update rule (detailing how the graph should be updated if the rule matches). Further to these, a rule can also make use of external

[4] http://www.sourceforge.net/projects/t-rex

gazetteer lists for reinforcing its precision by detecting better matches within a concept. The full flexibility of Saxon lies however in the ability to specify unrestricted Java code as the right hand side of a rule. The output of a rule can be either other annotations or unrestricted actions specified within the rule.

3.3 Corpus Collection

Online photostore users were contacted[5] for permission to use their public images to build a corpus of image metadata. During this period there were a total of 414 responses, of which 391 replied positively.

The corpus gathered for development and evaluation of our approach includes over 1.8 billion tokens distributed among over 119K image descriptions as it stands. This is largely characterised by short disconnected snippets of text (see fig.1) describing users photographs. In collecting the corpus, foreign language descriptions were inadvertantly collected and some minimal language filtering needed to be performed.

Language Filtering. In order to filter out foriegn language a scoring method based on the most common terms[6] of the British National Corpus (BNC) was devised. The idea was to reward the use of tokens within the annotations that are within this set of terms from the BNC and penalising the use of tokens that are not. This returns an estimate of the likelihood of any description being English and therefore included.

Training Data. A total of 1660 English image descriptions (24,215 tokens) belonging to 54 distinct users were randomly collected from the main corpus. This smaller corpus was then manually annotated by a group of 7 researchers according to the concepts introduced in section 2 generating a total of 2522 annotations. More specifically, 566 annotations were assigned to the concept of *Person*, 747 to *Location* and 1209 to *Object*. This dataset was then subdivided into 2 sets: the annotated data used by T-Rex as training data and Saxon as a basis from which to build extraction rules (40%), the remaining data was used for testing. Further image descriptions were also collected from the main corpus at a later stage for evaluating the approach.

Rule Development. The development of rules was an iterative process and took part in 3 stages: one for each concept defined in section 2. At the end of the process, 15 generic rules were developed. Four rules for *'Person'* aided by the use of gazetteer lists for detecting common first names, references to family relatives (e.g. mom, dad, brother, etc.) as well as person titles (e.g. Mr, Dr, etc). Six rules for *'Location'*, 5 of which were reinforced by gazetters for detecting common locations (e.g. countries, cities, etc.) as well as tokens indicative of references to a location (e.g. museum, street, etc). Five rules for *'Object'* were extracted, 3 were reinforced by organisation gazetteers to detect instances that refer to branded objects (e.g. McDonald's sandwich, Lincoln engine, etc).

3.4 Hybrid IE

In order to successfully extract information from image descriptions, it is argueable that either technique implemented by T-Rex or Saxon could be singularly applied to the

[5] 2325 Flickr users over a period of 4 months where contacted

[6] with frequency greater than 800

task. However, because of the constraints imposed by the domain and the requirements introduced in the beginning of section 3, each system carries with it disadvantages.

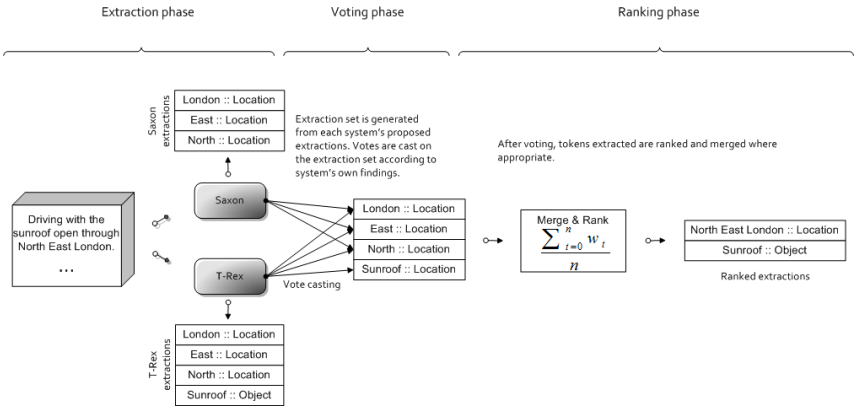
Despite implementing a flexible approach for IE, T-Rex depends heavily upon the size and coverage of the initial training dataset which is costly to develop. Also, when configured for performing highly accurate extractions computational cost can be impractical for use in scalable applications. Saxon on the other hand, while being less computationally expensive, requires time consuming development of rules for capturing every desirable case within the text, which makes it less flexible for performing IE. What we propose here is that the combining of the two techniques implemented by T-Rex and Saxon not only lessens their disadvantages, but also gives way to improved precision and recall while maintaining the approach as scalable as possible.

One of the first issues to be addressed by the combination of the two techniques is a architectural one. Machine learning approaches, as mentioned previously, utilise tokens' textual features from training data to build a generic data model that can be applied in previously unseen cases. In order for this data model to be highly accurate, multiple features must be recorded about as many neighbouring tokens as possible implying complexity and increased computational cost for an extraction task.

The domain of image descriptions as discussed previously is unique. Meaning for typically short texts, the size of the context a token can be placed in almost always shrinks down to 1 or 2 neighbouring tokens. The creation of a machine learning data model that reflects this reduces overall computational costs. On the other hand, in reducing the size of the contextual information gathered for the creation of an appropriate data model, the accuracy of extractions performed by T-Rex are also decreased. It now takes less constraints to be satisfied for finding a token fiting the model created. A potential solution would be to produce a greatly expanded training dataset but this would be a prohibitive option since it would not only be costly but also difficult to obtain a dataset that is comprehensive enough. The most suitable solution for improving the accuracy of extractions could therefore lie in the use of a rule based extractor.

Unlike in singularly developed rule based extractors, in developing a hybrid approach to IE, Saxon rules can be built in a generic way, thus speeding up development (i.e., less rules), as well as improving recall. While this would have a massive effect on precision in exclusively rule based systems, in a hybrid approach extractions can be compared according to different resources, thus giving rise to improved precision.

The essence of the approach therefore lies in extracting information from an annotation using a combination of the extraction suggestions from each system. So in order to better combine these extractions, a *weighted voting strategy* was devised that gave rise to an opportunity for taking advantage of both systems' strengths while attenuating the effects of their weaknesses. This *voting* method can be subdivided into three distinct phases: 1) **Extraction**: each system puts forward potential extractions found in an image description. 2) **Voting**: based on their separate findings, Saxon and T-Rex *"vote"* on each extraction according to a set of weights attributed to each system. 3) **Ranking**: the number of votes cast on the tokens of each extraction are used to give it a *"confidence"* ranking according to pre-specified ranges (i.e., between 0.8 and 1 - *high*, between 0.5 and 0.79 - *medium* or between 0 and 0.49 - *low*).



In the *extraction* phase, an image description is passed to each system separately and both generate a list of potential extractions from the original text together with their corresponding classifications (i.e., person, location or object). Once the potential extractions are identified, systems vote on the set of extractions based on their own findings and pre-defined weights. An obvious example of this would be in the description *"Driving with the sunroof open in North East London."* whereby both T-Rex and Saxon vote for all the tokens within *"North East London"* as referring to a location and only T-Rex votes for the token *"sunroof"* as referring to an object.

The set of votes V_t each token t receives can then be represented as $V_t = \{w_0, ..., w_r\}$ where w_r is the weight of the vote received from resource r (i.e. Saxon or T-Rex). The accumulated weight w_t for each token is obtained from the sum of the vote weights w_r that make up V_t for token t, see equation 1. The confidence ranking for each merged extraction E that is composed of n tokens where $E = \{t_0, ..., t_n\}$ can be obtained from the sum of each tokens' accumulated weight w_t divided by the number of tokens n that compose the extraction, see equation 2.

$$w_t = \sum_{w_r \in V_t} w_r \qquad (1) \qquad \frac{\sum_{t=0}^{n} w_t}{n} \qquad (2)$$

As exemplified, the votes are cast at the level of tokens. This allows extractions to be ranked according to what T-Rex and Saxon find regarding each single token that may be part of a larger entity. Once the accumulated weights for tokens are obtained, neighbouring tokens are then merged according to a combination of their weight, their extraction type and the confidence ranking expected from each extraction (i.e. *high*, *medium* or *low*). So in the example above, the tokens *North*, *East* and *London* are merged since their overall confidence ranking is very high (i.e. 1) and they were classified with the same type. However not all extraction combination scenarios are complimentary.

One of the strengths of this strategy is its ability to resolve overlapping extractions according to the three levels of confidence mentioned previously. A typical example is *"Autumn in Arlington cemetery"* whereby T-Rex extracts the token *Arlington* as a location and Saxon extracts *Arlington cemetery*. Both extractions are conceptually correct although one is more complete than the other. After voting the token *Arlington* would

arise as being a **high confidence** extraction, whereas the token *cemetery* would be classified as **medim confidence**. Depending on the confidence ranking expected, the final result could either be an extraction ranked with **medium confidence** that incorporates both tokens *Arlington cemetery* or an extraction ranked with **high confidence** that only includes the token *Arlington*. This is one of the advantages of using a *weighted voting strategy* in that it enables not only decisions on which extractions are the strongest, but also consider the ones that are not so strong as opposed to simply discarding them. One feature that arises from the existence of such rankings is that they allow the final extractions to be geared towards either one of high precision or high recall.

More problematic conflicts such as the disagreement regarding an extraction classification cannot be resolved by simply applying the three levels of confidence introduced above. This is where the full flexibility of a weighted voting strategy lies, in that the assigning of weights to votes can not only be used for ranking extractions but, when tweaked to reflect a higher confidence in the more precise technique at hand, can be used for resolving extraction type disputes across systems. An example found during the testing of the approach that would fit into this situation come from descriptions such as *"Auray in Brittany; North-West France"*, where *Auray* is clasified by T-Rex as a person and a location by Saxon. It is clear in this instance Saxon has classified the extractions correctly and this can be mainly attributed to the tokens being a correct match to an existing rule for extracting locations that is reinforced by a gazetteer list, thus yielding more precise extractions. Therefore in order to resolve conflict as exemplified above, the same voting strategy is used, but with the weights reflecting a higher confidence in Saxon as being the more precise technique in such circumstances and providing a means to resolve problems previously presented to either an exact match combination or an overlapping extraction. In the sections to follow, we present evaluation results obtained from this approach on an annotated subset of the main corpus and introduce possible future work.

4 Evaluation

So the evaluation of the task involved the detection of all occurances of locations, people and objects in an image description. The definition of how we decide whether extractions made are correct or not is crucial for the computation of evaluation scores. For the evaluation of the hybrid approach detailed earlier three different possibilities were considered: 1) **exact rule**: a prediction is only correct, if it is exactly equal to an answer. 2) **contain rule**: a prediction is correct, if it contains an answer, plus possibly a few extra neighboring tokens. 3) **overlap rule**: a prediction is correct, if it contains a part of a correct instance, plus possibly some extra neighboring tokens. An evaluation set of 100 previously unseen image descriptions that spanned the collections of 3 different users was randomly selected from the main corpus. This set was then manually annotated before being processed both by T-Rex and Saxon individually and as part of a hybrid system. The following results were obtained for extractions that were ranked in the *high* and *medium* confidence ranges.

	T-Rex		Saxon		Hybrid	
Concept	Precision	Recall	Precision	Recall	Precision	Recall
Person	67%	63%	70%	76%	86%	82%
Location	80%	62%	91%	77%	92%	79%
Object	75%	61%	75%	60%	73%	63%

As it can be seen from the results above, the hybrid approach outperforms T-Rex and Saxon when run individually for extracting instances of *'Person'* and *'Location'* from image descriptions, while for instances of *'Object'* there is no noticeable overall improvement. Each system's extractions were then shaped by their strengths and weaknesses and in most cases combined with great success using the hybrid approach. For instance, T-Rex was able to contextually detect the uses of unknown words to refer to locations depicted in the photograph such as *La Louvre* in *"Floaton on a fountain by La Louvre"* where Saxon failed. On the other hand, the usefulness of gazetteers and the precision of rules allowed Saxon to detect tokens such as *Harry Potter* in *"Of Harry Potter fame"* while T-Rex failed to do so.

Classification types can be corrected, in *"Low cloud on Mont Victoire"* T-Rex misclassified *Mont Victoire* as a person and Saxon correctly resolved the entity to a location. Other examples such as *"Big French sandwich"* and *"Worst seat in the best court"* demonstrated the flexibility of Saxon rules in complementing T-Rex's extractions of *sandwich* and *seat* with *Big French sandwich* and *Worst seat* which undoubtedly represent better conceptual extractions.

Although the approach performed well in combining both techniques there where some cases of misclassification. In most cases these occurred due to overgeneralisation both of Saxon rules and the T-Rex data model. Instances such as *life* in *"I have never seen anything like this in my life"* and *whole new meaning* in *"A whole new meaning for drive through"* were wrongly extracted as objects by either Saxon or T-Rex or a combination of both at times. Further to this, occasional entities such as *lines* in *"The people were lines up like crazy to get into this place"* cluttered the extraction set without adding any semantic value to it.

The issue of useful instances being overlooked by both systems can be partly attributed to part of speech misclassifications in descriptions such as *"Artwork! Sculptures in the sea at Crosby"*, *"Lifeboat on car ferry to France"*. In all descriptions, the references to objects of relevance within the photo (i.e. artwork, sculptures and lifeboat) are contextually difficult to be classified as objects (i.e. nouns instead of proper nouns), since their linguistic context also lends itself to other interpretations.

Finally, cases where there isn't enough linguistic content for performing extractions using only machine learning and rules or a hybrid approach are exemplified by descriptions such as *"Mull"* and *"French Riviera"*. Unless such noun phrases were already part of a pre-compiled gazetteer, the lack of a sentence structure surrounding such examples makes it very difficult to tackle the IE problem from a purely NLP perspective.

5 Conclusion and Future Work

In this paper, we have detailed a hybrid approach for extracting information from image descriptions that takes advantage of the combined results produced by systems that

implement widely used techniques for IE. More specifically we considered the combination of T-Rex, a machine learning framework, and Saxon, a rule based extractor, for addressing issues of computational cost as well as precision and recall when extracting information from such short snippets of text. As seen in the evaluation results, the use of a hybrid approach for extracting information from image descriptions is promising, however levels of precision and recall could be improved by using external knowledge for reinforcing the extractions. For instance, cases such as in the description *"Highland near Ben Nevis"* could be placed in the context of the user (e.g. does s/he know anyone called "Ben Nevis"?), the image itself (e.g. GPS positioning) or other image descriptions within the same collection (e.g. "Ben Nevis" was previously classified as a location/person). Another possible refinement to the approach, that has been previously applied with success in the past for the task of image annotations [1], is that of involving the user in the process for reinforcing system decisions, such as confirming the outcome of a conflict resolution.

Furthermore, the concepts used here are an incomplete list of those useful within an image description. One important area for future work is extraction of further concepts used by people to describe their images (e.g. time, events, mood, etc). Also, some extraction examples, such as in the description *"Vicky and dad at local bus stop"* where *local bus stop* is extracted as an object, suggest that certain concepts may need further refinement. This would allow in this case for the object instance found to be also assigned geographic properties, given the contextual information about the image.

Acknowledgements This work was sponsored by Kodak Eastman Corporation. We would also like to thanks the 391 online photo sharing users who donated their photos and captions.

References

1. L. Ahn and L. Dabbish. Labeling images with a computer game. In *CHI '04*, pages 319–326, New York, NY, USA, 2004. ACM Press.
2. A. Barla, F. Odone, and A. Verri. Old fashioned state-of-the-art image classification. In *Proc. of ICIAP 2003*, pages 566–571, Sept 2003.
3. M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proc. of COLING 1992*, pages 539–545, 1992.
4. J. Iria and F. Ciravegna. A Methodology and Tool for Representing Language Resources for Information Extraction. In *Proc. of LREC 2006*, Genoa, Italy, May 2006.
5. J. Iria, N. Ireson, and F. Ciravegna. An experimental study on boundary classification algorithms for information extraction using svm. In *Proc. of EACL 2006*, April 2006.
6. M. Naaman, S. Harada, Q. Wang, H. Garcia-Molina, and A. Paepcke. Context data in georeferenced digital photo collections. In *Proc. of ACM MM*, Oct 2004.
7. K. Pastra, H. Saggion H, and Y. Wilks. Extracting relational facts for indexing and retrieval of crime-scene photographs, 2002.
8. R. Srihari. Automatic indexing and content-based retrieval of captioned images. *Computer*, 28(9):49–56, 1995.
9. R. Veltkamp and M. Tanase. Content-based image retrieval systems: A survey. Technical Report UU-CS-2000-34, Dept. of Computing Science, Utrecht University, 2000.