Automatic Image Captioning From the Web For GPS Photographs

Xin Fan[1], Ahmet Aker[2], Martin Tomko[3], Philip Smart[4], Mark Sanderson[1], Robert Gaizauskas[2]

[1]Department of Information Studies, University of Sheffield, UK

[2]Department of Computer Science, University of Sheffield, UK

[3]Department of Geography, University of Zurich, Switzerland

[4]School of Computer Science, Cardiff University, UK

x.fan, a.aker, m.sanderson, r.gaizauskas@shef.ac.uk, martin.tomko@geo.uzh.ch,
p.smart@cs.cf.ac.uk

ABSTRACT

Increasing quantities of images are indexed by GPS coordinates. However, it is difficult to search within such pictures. In this paper, we propose a solution to automatically generate captions (including place name, keywords and summary) from the web content based on image location information. The richer descriptions have great potential to help image organisation, indexing and search. The solution is realised through the synergetic techniques from Geographic Information System, Web IR and multi-document summarisation.

Categories and Subject Descriptors

H.3.1 **[Information Storage and Retrieval]:** Content Analysis and Indexing – *Abstracting methods, indexing methods;*

General Terms

Algorithms, Human Factors.

Keywords

Image captioning, image search, geographic information system, term weighting, concept ontology, multi-document summarisation.

1. INTRODUCTION

Images with location information are growing rapidly after the introduction of location indexing in photo sharing services, such as Flickr and Panoramio, and with increasing prevalence of cameras with embedded GPS. However, creation of text captions based on a location is only possible through manual annotation. As such annotation is time-consuming and laborious, many efforts have been made on automatic generation of image tags and captions. If images are embedded in documents, image captions can be automatically generated from existing image captions, surrounding text etc. [4].

However, most of the existing research on image captioning focuses on the "*what*" aspect of an image and try to discover *what* is in the image. In this paper, we propose an automatic image captioning solution to provide semantic descriptions based on image location derived from. We mainly focus on captioning such

photos in order to make these photos searchable. In contrast to previous image captioning work, our solution is targeting the "*where*" aspect of an image and exploring the location-based knowledge for automatic image captioning.

The augmented caption consists of generated keywords, a paragraph of readable text summary and an identified place name. We take advantage of spatial data from a Geographic Information System (GIS) to derive semantic informatiodn about the image location. Then the location-based information is effectively combined with techniques from Web IR and text summarisation to generate image captions.

The enriched and readable captions can help image search and browsing in many ways. The generated image caption can be used as text features in image indexing to make the photos only associated with coordinates searchable in an image search engine. In addition, the current recommendation of GPS photos is mostly only based on location proximity. The generated captions can enable image recommendation services to link the photos in terms of both spatial and semantic connections.

In our solution, an image caption consists of an identified *place name*, a set of *concept keywords*, a set of *expanded web keywords* and a text paragraph of *summary* of the identified place.

A photo associated with coordinates is geo-referenced into a corresponding toponym (*place name*) and a footprint of the captured view.

A GIS interrogates a digital map (spatial data) to determine the land cover type(s) in the footprint and this is mapped into a possible *scene type* of the photo [7]. We introduce location based concepts from a geographic concept ontology [5] in terms of the assigned *scene type*. Furthermore, we use the statistics from web search result to rank the keywords and filter out the less relevant ones in respect of the captured place (represented by a toponym). The top ranked terms are selected as *concept keywords*.

We use the place name and concept keywords as query to retrieve relevant web pages. The text segment related to the place name and concept keywords are collected and form a text paragraph. Furthermore, the *expanded web keywords* are extracted from the text paragraph.

A short *summary* is generated by a language model based extractive multi-document summariser. The summariser uses the identified toponym to retrieve related web pages and applies language models to select salient sentences about the image

location. The extracted sentences are organised together into a summary.

2. Image Geo-referencing

Image geo-referencing is a process of determining the spatial location and other geographical features that can be analysed (such as map projections, footprint etc.) for an image.

As describe in Section 1, the image coordinates are used to identify a place name of the captured view and produce a geographic footprint of the captured view, which is then related to a place name through a toponym ontology [1].

In order to identify the *place name*, the image location is classified into a rural or urban area determined by querying spatial data for land cover types contained within the footprint, and their classification into rural or urban classes. A circular section area is defined centred at the image coordinates and different buffer sizes are assigned in terms of rural or urban image types. Images with no directional information are assigned a circular buffer. All the candidate toponyms were ranked by a salience value calculated by the geographical distances and web occurrences of the toponyms. The details are described in the work [1, 7]. The highest ranked toponym was identified as the corresponding place name for the image location. As shown in Fig. 1, the image with the GPS coordinates will be mapped to a toponym "Westminster Abbey".

Embedded EXIF:
GPS Latitude Ref = North latitude
GPS Latitude = 51.499532
GPS Longitude Ref = West longitude
GPS Longitude = -0.12913

Toponym: Westminster Abbey

Parent Toponym: London, England, GB, United Kingdom

Circular Footprint (based on GPS coordinates, without orientation information)
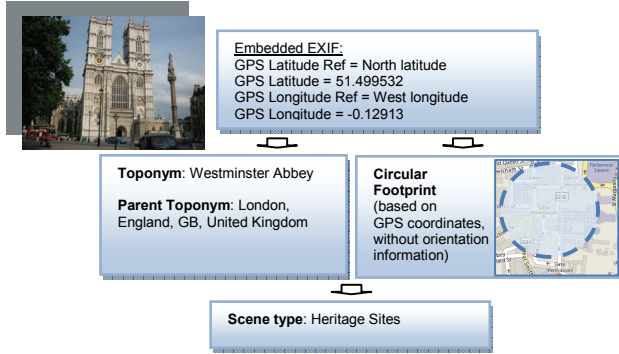
Scene type: Heritage Sites

Fig. 1 Illustration of image geo-referencing process

Based on the footprint of the captured view, we further explored and identified the common *scene type*s for an image. The *scene type* of a certain image location is derived from geographic feature type in the map data and the land cover type. For example, we assigned the *scene type* "Historic Sites" to the geo-referenced image in Fig. 1. According to the image *scene type*, a number of distinct concept terms were extracted from a geographic concept ontology [5]. The concept terms describe the *scene type* from different aspects. In addition, the *scene type* is fed into the Multi-Document Summariser (MDS) in Section 5.

3. Geographic Concept Weighting and Filtering

In this section, we briefly introduce how a set of concept terms associated with a *scene type* are generated, which can be regarded as a complete term set describing a *scene type*. Furthermore, we rank the terms based on the weight scores and select a subset of most probably relevant terms for a specific *place name* (toponym of the geo-referenced image) from the term set. The selected terms will be used as image *concept keywords* and the input for extracting *expanded web keywords* and *summary* related to the image location.

We generate a set of related terms $T_S\{t_1, t_2 \cdots t_n\}$ to a certain *scene type* S based on an geographic concept ontology [5], in which the vocabularies generally describing a *scene type* are organised in a term list $\{t_1, t_2 \cdots t_n\}$

Therefore, we regard the concept terms $T_S\{t_1, t_2 \cdots t_n\}$ as a complete set to describe the *scene type* S. We select a sub-set of terms $T_{Topo}\{t_1, t_2 \cdots t_m\}$ for a specific place *Topo* (the corresponding toponym) belonging to the *scene type* S.

$$T_{Topo}\{t_1, t_2 \cdots t_m\} \subset T_S\{t_1, t_2 \cdots t_n\} \quad (1)$$
$$m < n, \; Topo \; with \; Scene \; Type \; S$$

The toponym dependent terms $T_{Topo}\{t_1, t_2 \cdots t_m\}$ are selected in a process of weighting, ranking and filtering of the *scene type* terms $T_S\{t_1, t_2 \cdots t_n\}$.

Thus, we first estimate the weight $W_t(t_i | Topo)$ of a term $t_i \in T_S\{t_1, t_2 \cdots t_n\}$ in respect of a specific place named *Topo* by a mix of features including term-SceneType weight $W_{t-s}(t_i | S)$, term-toponym weight $W_{t-t}(t_i | Topo)$ and term popularity $P_t(t_i)$.

$$W_t(t_i | Topo) = \frac{W_{t-t}(t_i | Topo) / P_t(t_i)}{\sum_{t_i \in T_S} W_{t-t}(t_i | Topo) / P_t(t_i)} \quad (2)$$

where $W_{t-t}(t_i | Topo)$ is the term weight in respect of a certain place name *Topo*. It is estimated based on the co-occurrence rate t_i and *Topo* in the Yahoo web search result. $P_t(t_i)$ measures the popularity of the term t_i. It is calculated as $P_t(t_i) = \sqrt{tf_{Web}(t_i)}$ where $tf_{Web}(t_i)$ is the term occurrence frequency in the Yahoo! web search result.

The concept terms $T_S\{t_1, t_2 \cdots t_n\}$ are ranked by their weights calculated in the equation (2) and sorted in a descending order. Since the term weight is normalised to a value within (0, 1), we choose the terms by the following strategy. First, the concept terms $T_S\{t_1, t_2 \cdots t_n\}$ are sorted to $T_S\{t_1', t_2' \cdots t_n'\}$ by their weights in a descending order. We then add the terms in the order until the sum of their weights are larger than a threshold θ.

$$T_{Topo}\{t_1, t_2 \cdots t_m\} = T_S\{t_1', t_2' \cdots t_m'\},$$
$$\sum_{i=1}^{m} W_t(t_i' | Topo) \geq \theta, \sum_{i=1}^{m-1} W_t(t_i' | Topo) < \theta \quad (3)$$

The top concept terms $T_{Topo}\{t_1, t_2 \cdots t_m\}$ are regarded as *concept keywords* in the generated image caption.

4. Web Keyword Expansion

Most of *concept keywords* are single words and describe the capturing location in a geographic context. In order to discover the location descriptions from different aspects, the *expanded web keywords* are extracted from related web pages in terms of place name and *concept keywords*, which can be single words or phrases. The web keyword expansion includes the following four steps: (1) descriptor identification; (2) web document retrieval and content extraction by the descriptors; (3) window text (neighbour text segments) extraction in terms of descriptors; (4) keyword extraction from the collection of window texts.

4.1 Descriptor identification

We adopt the standard place name (e.g. "Westminster Abbey") in the identified toponym hierarchy (e.g. "Westminster Abbey, London, England, GB, United Kingdom") and the *concept keywords* $T_{Topo}\{t_1, t_2 \cdots t_m\}$ as the descriptors.

4.2 Web document retrieval and content extraction

The web query is formulated by combining the descriptors of the standard place name $topo_{standard}$ and the *concept keywords*. The upper-level toponym $topo_{parent}$ (for example, "London" in the above toponym hierarchy) is also added in order to disambiguate the descriptors. A faceted query is in a format as: $topo_{standard}$ AND $topo_{parent}$ AND $(t_1 \text{ OR } t_2 \text{ OR } \cdots \text{ OR } t_m)$. The query is sent to Yahoo! search API to collect the top 30 web pages. The elements in the pages are filtered to remove the superfluous content such as navigations, ads, scripts etc. by traversing the DOM tree nodes.

4.3 Window text extraction

Considering the most relevant text information to a descriptor usually exists in the neighbouring words within the same sentence, we use a fixed-size sliding window within a sentence to collect relevant text segments. For instance, the word "Westminster Abbey" is an identified descriptor, a maximum 30-word window is employed in the below text paragraph to select the underlined relevant text segment in the same sentence.

```
Every year Westminster Abbey welcomes over one
million visitors who want to explore this
wonderful 700-year-old building which is the
coronation church of England. Thousands more flock
to the Abbey for worship at daily services.
```

The text segments for the identified descriptors in a specific web page are collected to form a *virtual document d*. All the *virtual documents* are included a document collection $D_{Web} = \{d\}$ which describes the image regarding the geographic location. The web keywords are extracted from this document collection.

4.4 Keyword extraction

Each text segment in a *virtual document d* is first split into a set of simple words by a tokeniser. Here "word" is defined as any combination of letters, including hyphens, but excluding symbols and punctuation, e.g. apostrophes. The tokenised text is tagged by a POS (Part of Speech) Tagger afterwards.

All the word sequences containing 1-4 words within a text segment are listed as candidate phrases. The candidate phrases cannot start with or end with a stop word. The stop word list contains the following words which can be annotated in the POS tagger: coordinating conjunction, determiner, preposition or subordinating conjunction and wh-determiner. As the noun phrases are generally most informative for the image location, we only keep the phrases ending with a noun or a proper noun in the candidate phrase list.

The candidate phrases are also validated by a cross-page occurrence requirement. Phrases that occur in less than 2 *virtual documents* (i.e. occur in less than 2 different web pages) are removed from the candidate list.

The score of a candidate phrase p is calculated as the following equation to rank the candidate phrases:

$$S(p) = P_{Web}(p) - IDF_{VC}(p) = \log \frac{|N_{Web}|}{|n_{Web}(p)|} - \log \frac{|D_{VC}|}{|d_{VC}(p)|} \quad (4)$$

where $|n_{Web}(p)|$ is the occurrence frequency of candidate phrase p in the Google n-gram corpus, $|N_{Web}|$ denotes the occurrence frequency of the most popular word "the" in the Google n-gram corpus. $|d_{VC}(p)|$ is the document frequency containing the candidate phrase p in the *virtual document* collection $D_{Web} = \{d\}$. $|D_{VC}|$ is the total number (document frequency) of the *virtual documents* in the collection D_{Web}.

All the candidate phrases are ranked according to the scores and the top ones are selected as the *expanded web keywords*.

5. Summary Generation

The summary paragraph for image caption is generated using *the-MDS* (the-multi-document summariser), an extractive, language independent, multi-document, query-based summarization system [2]. The inputs are the identified toponym and image *scene type* for each image as well as web-documents retrieved using the Yahoo! Search engine. The query to the search engine is the toponym. The summariser uses a three step approach to create image captions. It first applies shallow text analysis to the given web-documents. Then it uses a set of features to identify salient sentences. Finally, it performs sentence selection on the salient sentences to create the final summary.

5.1 Shallow text analysis

The summarizer first applies shallow text analysis including sentence detection, tokenization, lemmatization, named-entity recognition and POS-tagging to the given documents. For performing each of the pre-processing steps we use OpenNLP (http://opennlp.sourceforge.net/).

5.2 Feature extraction

After text analysis, *the-MDS* represents each sentence in the documents as a vector, where each vector position contains a term (word) and a value which is a product of the term frequency (TF) in the document and the inverse document frequency (IDF), a measurement of the term's distribution over the set of documents [6]. The IDF table is generated on the fly from the n related documents. The vector representation is used to calculate the following features for each sentence (using cosine similarity):

1. **querySimilarity**: Sentence similarity to the query/toponym.
2. **centroidSimilarity**: Similarity to the centroid. The centroid is composed by 100 words occurring most frequently in the document collection.

In addition to these features the summarizer adds to each sentence three further features:

3. **sentencePosition**: Position of the sentence within its document. The first sentence in the document gets the score 1 and the last one gets $1/n$ where n is the number of sentences in the document.
4. **starterSimilarity**: A sentence gets a binary score if it starts with the query term (e.g. Westminster Abbey, The Westminster Abbey, The Westminster or The Abbey) or with the scene or object type, e.g. The church.
5. **modelSimilarity**: As a fifth feature we use language models (Section 5.2.1) to bias the sentence selection. The score is calculated according to the following Formula:

$$\mathrm{mod}\,elScore(S,M) = \prod_{ngram \in S}(prob_{ngram}+1) \qquad (5)$$

The modelSimilarity score of a sentence S is the product of scores ($prob$) of its n-grams where the $prob$ values are obtained from the language model M (Section 5.2.1).

Finally, for each sentence these features are combined to a final score using linear regression. The final score is used to rank the sentences.

5.2.1 Scene type language models

Language models are used in different fields with different purposes. In information retrieval (IR), for instance, language models are used to retrieve documents relevant to a query. For each document a distinct n-gram language model is derived and used to estimate the probabilities of producing each term in the query [3]. The query is treated as a generation process, i.e. based on each language model the probability of generating each term in the query is computed. The probability of generating the query is the product of terms occurring in the query. Finally, the documents are ranked in descending order based on the probability assigned to the query. Therefore, if terms of a document lead to higher generation probabilities, the more relevant this document is to the query.

We also generated language models from Wikipedia articles about the same scene type [2]. Similar to [3], our language models are used in a generative way, i.e. we calculate the probability that a sentence is generated based on an n-gram language model.

5.3 Sentence selection

After the scoring process, the summarizer selects sentences for summary generation by selecting the sentence from the first position from the ranked list, followed by the next non-redundant sentence in the list until the summary length reaches 200 words [2]. Fig. 2 shows an example of the automatically generated caption for a photo of the Westminster Abbey.

Fig. 2 An example of generated caption

6. Conclusion

In this paper, we presented a solution for automatic captioning photographs associated with geographic coordinates. The image caption, including an identified *place name*, a set of *concept keywords*, a set of *expanded web keywords* and a text paragraph of *summary* of the identified place, were created using the synergetic techniques from Geographic Information System, Web IR and multi-document summarisation. In the future work, we will investigate the semantic relationships between generated keywords in keyword ranking and filtering. In addition, we plan to introduce generic sentence templates by information extraction techniques to produce image captions.

7. Acknowledgements

8. References

[1] Abdelmoty, A. I., Smart, P. and Jones, C. B. Building place ontologies for the semantic web: issues and approaches. *the 4th ACM workshop on Geographical information retrieval* (Lisbon, Portugal, Nov. 2007)

[2] Aker, A. and Gaizauskas, R. Summary Generation for Toponym-Referenced Images using Object Type Language Models. *the RANLP 2009* (2009)

[3] Croft., F. S. a. W. A general language model for information retrieval. *Proc. of the eighth international conference on Information and knowledge management* (New York, NY, 1999), 316-321.

[4] Deschacht, K. and Moens, M. F. Text Analysis for Automatic Image Annotation. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (Prague, Czech Republic, June 2007), 1000-1007.

[5] Purves, R. S., Edwardes, A. and Sanderson, M. Describing the where – improving image annotation and search through geography. *Proceedings of the workshop on Metadata Mining for Image Understanding* (Madeira, Portugal, Jan 2008)

[6] Salton, G. and Buckley, C. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24, 5 1988), 513-523.

[7] Tomko, M., Purves, R. S., Hall, M. M. and Smart, P. *Tripod Deliverable 6.2: Report and Prototype for Enriching Geo-referenced Images with Spatial Data*. EU FP6 Project no. 045335 TRIPOD (TRI-Partite multimedia Object Description), Jun. 2009.