Automated Image Captioning : The TRIPOD project

Christopher B. Jones, Alia I. Abdelmoty, Philip D. Smart, Mark M. Hall,
Jonathan Quinn and Florian A. Twaroch

School of Computer Science, Cardiff University, Cardiff, UK

1 Introduction

There is a vast number of images available on the web but effective retrieval facilities to find the images we want are limited by the fact that most of the images carry very little descriptive information. While image context analysis holds some promise to find images of certain types of phenomena, there remains the need to enrich images with machine and human readable metadata that can be used to match text-based queries. The TRIPOD project sets out to remedy current limitations by developing methods to improve the quality of metadata associated with stored images.

There are two primary strands to this research. The first is to generate data that can be associated with existing manually-captioned (archival) images. The second is to generate captions and metadata automatically for images that have no captions but which have been geo-located (for example with GPS) and which may also be oriented if the camera had a 2D or 3D compass. In both cases the focus of the generated metadata is on the geographical context of the images. A prototype image search engine is being constructed to demonstrate the results of the research.

TRIPOD is a multi-partner EC-funded project with academic and professional expertise. The academic partners are from the universities of Sheffield (UK), Zurich (Switzerland), Bamberg (Germany), Dublin City (Ireland) and Cardiff (UK) and have expertise in information science, computer science and geographical information science. The professional expertise covers the fields of image archive and search facilities, geo-data provision and software development from Alinari and Centrica in Florence (Italy), Geodan (Netherlands), Tilde (Latvia) and Ordnance Survey (UK).

2 Enhancing existing captioned images

For images that already have a caption, TRIPOD provides functionality to geo-reference the image and to generate keywords. The process of geo-referencing requires parsing the image caption to detect the presence of a toponym (place name) and then, if necessary, disambiguation of the toponym. Toponym detection and disambiguation are performed with the assistance of a toponym ontology that can access both local and remote multilingual gazetteer resources. The toponym ontology is being enhanced with knowledge of vernacular place name terminology derived from the web as well as with specialized language processing facilities to support less-widely used languages such as Latvian. In order to assist in interpretation of

spatial terminology of existing captions, and hence derive the footprint of an image, an empirical study of uses of spatial natural language is being conducted.

Several techniques and resources are exploited to generate additional concept terms or keywords that can be used to characterise the image with regard to the scene elements and qualities. The nature of the local landscape of the image can be inferred by using the footprint resulting from geo-referencing to retrieve data from thematic geo-spatial datasets. In particular, landcover resources such as the CORINE dataset have been employed. Further potentially relevant keywords can be found with a concept ontology that associates generic scene types with sets of related concepts. Some of the content of the concept ontology has been derived from studies of the terminology employed to caption images in a substantial geographically-referenced image database (http://www.geograph.org.uk) and in Fickr (http://www.flickr.com), as well as from human-subject studies of terminology used to describe images.

Potentially-relevant concept terms can be refined by searching the web with the toponym as a keyword to find which concepts are most commonly associated with that particular instance of a place type. In addition to generating sets of keywords the project is applying multi-document web summarisation methods to provide richer natural language descriptions of what is assumed to be visible in the image.

3 Generating captions for geo-located images

Methods to generate captions from geo-located images without pre-existing captions depend upon the nature of the associated locational data. If just a GPS coordinate is present then that coordinate can be used for reverse geo-coding to find a potentially relevant place name. The GPS location can also be used to infer the landscape type from geo-spatial data, while terrain models can be used to generate a 360^0 viewshed that indicates which nearby features could be visible in the image. If camera orientation data are also available, then more sophisticated approaches can be adopted. The azimuth and focal length can be used to narrow down the candidate visible objects and to suggest landmarks that may be expected to be visible. A major development within the project however is the exploitation of 3D city and rural models to infer with relatively high precision which objects are in the field of view, given knowledge of the 3D orientation and focal length of the camera. The visually and semantically salient features can then be selected for mention in image captions that employ spatial language based on the empirical studies. As with previously captioned images, multi-document summarisation techniques can again be used to provide richer descriptions of features known to be in the field of view.

Acknowledgements

This work has also been funded by the EC FP6-IST 045335 TRIPOD project. More information can be found on the website of the project: http://tripod.shef.ac.uk/