Title	Leveraging Ontologies, Context and Social Networks to Automate Photo Annotation
Author(s)	Monaghan, Fergal; O'Sullivan, David
Publication Date	2007
Publication Information	Fergal Monaghan, David O'Sullivan "Leveraging Ontologies, Context and Social Networks to Automate Photo Annotation", Proceedings of the 2nd International Conference on Semantics and Digital Media Technologies (SAMT 2007), 2007.
Publisher	Springer
Link to publisher's version	http://dx.doi.org/10.1007/978-3-540-77051-0
Item record	http://hdl.handle.net/10379/418
DOI	http://dx.doi.org/10.1007/978-3-540-77051-0

Leveraging Ontologies, Context and Social Networks to Automate Photo Annotation

Fergal Monaghan and David O'Sullivan

Digital Enterprise Research Institute,
National University of Ireland, Galway,
IDA Business Park, Lower Dangan,
Galway, Ireland
fergal.monaghan@deri.org,david.osullivan@nuigalway.ie
http://www.deri.org

Abstract. This paper presents an approach to semi-automate photo annotation. Instead of using content-recognition techniques this approach leverages context information available at the scene of the photo such as time and location in combination with existing photo annotations to provide suggestions to the user. An algorithm exploits a number of technologies including Global Positioning System (GPS), Semantic Web, Web services and Online Social Networks, considering all information and making a best-effort attempt to suggest both people and places depicted in the photo. The user then selects which of the suggestions are correct to annotate the photo. This process accelerates the photo annotation process dramatically which in turn aids photo search for a wide range of query tools that currently trawl the millions of photos on the Web.

Key words: Photo annotation, context-aware, Semantic Web

1 Introduction

Finding photos is now a major activity for Web users, but they suffer from information overload as their attempts to find the photos they want are frustrated by the enormous and increasing number of photos [1][2]. The processing speed of computers can be leveraged to perform the hard work of finding photos for the user and to thereby alleviate the information overload. To enable the machine to retrieve photos for the user, we must first examine how users mentally recall photos themselves. Research indicates that users recall photos primarily by cues in the following categories, in descending order of importance: (i) who is in the photo; (ii) where the photo was taken; and (iii) what event the photo covers [3]. Searchable description metadata in these categories can be created for photos and search engines can then match user queries with these descriptions and present the best matches to the user. A key challenge is how to create this useful, searchable description metadata about photos. Manual annotation of photos is tedious and consumes large amounts of time. Automated content-based techniques such as face recognition rely on large training sets and are dependant

on the illumination conditions at the scene of photo capture [4]. Complimentary context-based approaches provide a lightweight, scalable solution to support the abstract way in which users actually think about photos. Section 2 introduces the implementation of just such a context-based approach: the Annotation CReatiON for Your Media (ACRONYM) prototype[1].

Related work. CONFOTO [5] is a semantic browsing and annotation service for conference photos. It combines the flexibility of the Resource Description Framework (RDF) with recent Web trends such as folksonomies, interactive user interfaces, and syndication of news feeds.

PhotoCompas [6] uses timestamps and co-ordinates captured by GPS-enabled cameras to lookup higher level contextual metadata about photos from existing Web services. Given metadata from previously annotated photos it suggests people that may be depicted in consequent photos.

ZoneTag[2] is a prototype for Nokia S60 smartphones that allows the user to upload images from the phone to the Flickr website. Zonetag leverages the context (e.g. location and time) captured by the smartphone to find a location tag and to suggest other Flickr tags based on tags previously entered by the user and their social network under a similiar context.

2   Annotation CReatiON for Your Media

ACRONYM is a Semantic Web-based photo annotation tool that can annotate any JPEG image on the Web with RDF. It focusses on the most important recall cues and takes advantage of RDF's powerful expressivity, interoperability and mobility while hiding its complexities from the user. ACRONYM makes use of the EXchangeable Image File (EXIF) format metadata that is created and stored inside JPEG photo files by off-the-shelf digital cameras. This commonly includes a timestamp of when the photo was captured, shutter speed, exposure time etc. but can also include the co-ordinates of the camera at the time of capture if the camera has been coupled with a GPS receiver. ACRONYM also makes frequent use of the GeoNames[3] geographical database, map, ontology and Web services.

The user logs in with their email address: this is hashed to provide a unique identifier and to link the user to any RDF metadata describing them. Once logged in the user can add people, places and import arbitrary RDF to the system with the click of a button. The user selects which JPEG photo to annotate by specifying its URL. The system displays the JPEG image and translates its EXIF metadata into RDF metadata formalised in a combination of the Dublin Core Terms, Friend-of-a-Friend (FOAF), World Geodetic System 1984 positioning and GeoNames ontologies. This RDF is then combined with similar metadata from other photos to provide suggestions to the user for the creation of further metadata about the photo.

---

[1] `http://acronym.deri.org`

[2] `http://zonetag.research.yahoo.com`

[3] `http://www.geonames.org`

The user selects the people depicted in the photo from a list of suggested candidates that are described by FOAF social network metadata within the system. FOAF metadata about people and the relationships between them and the user can be created at the click of a button or imported from external sources. Instead of trying to identify faces in the content of the image, ACRONYM analyses the social context of the photo, ranking and ordering candidates in the list based on their social connection to the photo and the user as described by the `<foaf:knows>` relationship. A candidate receives one ranking point for each direction of a `knows` relationship between them and the user. Once at least one person has been selected as being depicted in the photo, an additional metric is used: one ranking point is added to each candidate per `knows` relationship between them and each person depicted. As the user selects which of the candidates are depicted, the list of candidates is updated: those people in the social circle of numerous depicted people float to the top of the list. This captures the social context of the photo and user and makes effective use of it by homing in on the most likely people to be annotated as depicted.

RDF metadata about places can be added via a full-text search field that queries a GeoNames Web service or can alternatively be imported from external sources. The user can also import from GeoNames all metadata on places within a specified kilometre radius of a selected place. Similiar to above the user then selects the places depicted in the photo from a list of suggested candidates in the system. If co-ordinates have been supplied by the EXIF metadata, these are used to lookup and import metadata about nearby places from GeoNames.

If (as is the common case) no co-ordinates have been supplied, ACRONYM again takes an analytical approach to estimate where the photo was captured. The algorithm takes the set of people already annotated as being depicted in the photo and estimates the location of each person at the time the photo was captured based on the co-ordinates of places they are co-depicted with in previously annotated photos. Firstly, each person is analysed to determine the temporally closest, previously annotated photo that depicts them. The mean of the co-ordinates of the places depicted in this other photo provides an estimate of where that person was at that time. The timestamped co-ordinates estimate for each person is then used to obtain weighted mean co-ordinates (weighted by temporal proximity to the photo being annotated) as a rough estimate of where all the people are co-depicted in this photo. This rough estimate is then used in place of hard data captured by a GPS receiver.

Each place in the system is then ranked according to its geographic proximity to where the photo was captured (or was estimated to have been captured) and this ranking is used to order the suggested places list for the user to select actual depicted places from. Once there is at least one place selected as being depicted, the mean co-ordinates of each depicted place are used in place of the rough estimate above to rank the suggested places. As the user selects which candidates are depicted, the list of candidates is updated: those places nearby numerous depicted places float to the top of the list. This captures the geographic

context of the photo and makes effective use of it by homing in on the most likely places to be co-depicted with the given people and places at the given time.

3   Future Work

The main thrust of future work is to integrate event suggestion with that of people and places. Cluster analysis will be implemented on the temporal, geographic and social aspects of photos to detect abstract events which can be concretely named, suggested and annotated to photos. A key concept of future efforts will be two slider bars to tune precision and recall: only candidates with a ranking that meets the recall setting will be suggested and those suggested candidates that also meet the precision setting will be automatically annotated to the photo. The user will be able to quickly tune the automation level from fully manual to fully automatic. Furthermore, readily available face detection tools will be assessed to locate and count people depicted in photos.

4   Conclusions

ACRONYM's suggestion algorithm captures and makes use of key context cues. By looking up existing information and inferring higher level contextual knowledge the tool accelerates the photo annotation process. The end result is that from ground truth machine-readable metadata captured by cameras, such as time and co-ordinates, ACRONYM discovers human-readable fields such as placenames and people names that are actually useful to query engine users. This alleviates the information overload on users searching for photos on the Web.

References

1. Infotrends: Worldwide camera phone sales to reach nearly 150 million in 2004, capturing 29 billion digital images. Technical report (2004)
2. Infotrends: Worldwide consumer digital camera sales to reach nearly 53 million in 2004. Technical report (2003)
3. Mor Naaman, Susumu Harada, Q.W.H.G.M.A.P.: Context data in georeferenced digital photo collections. In: 12th International Conference on Multimedia (MM 2004), New York, NY (2004)
4. Alice J. O'Toole, P. Jonathon Phillips, F.J.J.A.N.P.H.A.: Face recognition algorithms surpass humans matching faces over changes in illumination. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(9) (2007) 1642–1646
5. Nowack, B.: Confoto: Browsing and annotating conference photos on the semantic web. Web Semantics: Science, Services, and Agents on the World Wide Web **4**(4) (2006) 263–266
6. Naaman, M.: Leveraging Geo-Referenced Digital Photographs. PhD thesis, Stanford University (2005)