Department of
Information Engineering
and Computer Science DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

CONTEXT-BASED MEDIA GEOTAGGING OF PERSONAL PHOTOS

Ivan Tankoyeu, Julian Stöttinger, Fausto Giunchiglia

March 2013

Technical Report # DISI-13-018

Context-based Media Geotagging of Personal Photos

Ivan Tankoyeu, Julian Stöttinger, Fausto Giunchiglia
DISI, University of Trento
via Sommarive 14
38123 Povo, Trento, Italy
[tankoyeu | julian | fausto]@disi.unitn.it

ABSTRACT

This paper addresses the problem of automatic geotagging of media within the context of a personal media collection. In contrast with textual and visual methods which tackle the same problem we approach it focusing on analysis of contextual information. An event as a context aggregator plays the central role in our approach. The proposed method automatically estimates geographical coordinates (*latitude* and *longitude*) within the temporal boundaries of events computed from a personal media collection. Proposed framework interpolates or extrapolates GPS information rely on geoannotated media entities from the collection. The process of interpolation is automatically performed by the framework based on temporal distances between samples in combination with using free on-line navigation service. All this leads to a new cost efficient and intelligible event-centered way to enrich the collection with geographical information. Experimental results show that we are able to assign geographical coordinates for 83% of images within an error of 5 km.

Categories and Subject Descriptors

H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval; G.1 [**NUMERICAL ANALYSIS**]: Interpolation

General Terms

Algorithms, Performance, Experimentation

Keywords

Media Geotagging, Personal Media Collection, Context Processing

1. INTRODUCTION

The widespread of GPS[1]-enabled digital cameras and camera phones leads to the increasing number of geo-annotated photos. The wide use of spatial information in multimedia is supported by photo management software and on-line sharing tools. Recent studies have shown the importance of geographical information to a user for organizing personal photo collection [15]. This unveils for a user the possibility of sorting and organizing one's digital media collection in geospatial modality. Moreover additional services can be provided based on spatial information extracted from personal media collection [3].

However the vast majority of photos and videos uploaded to on-line sharing services are not geotagged. If they are, the GPS information is not available for all images, or manual annotation is only done for a few images. Therefore automatic techniques for assigning geographical coordinates to the digital media are required [2]. Current state of the art techniques approach this problem using textual and visual analysis. Both techniques require prior training of classifier and availability of a training set for this task. All this leads to a decrease in efficiency. In contrast to the current state-of-the-art methods our approach analyses the context. By the context we mean the spatio-temporal information related to the image provenance. We claim that in the scope of the entire collection of an individual user, the spatio-temporal context information is at least as important for analysis as it is visual content.

The central thesis of our paper is to leverage personal events for the task of geo annotation. The importance of event-based indexing for personal photo collection have been recently studied in [1]. Events can be seen as useful entities that provide a way to encode contextual information, and aggregate media that constitute the experience of such event. Events being context aggregators bring semantically meaningful information for a user. Due to the nature of an event space and time information is the most important data to identify an event. However, time information is the primary attribute for detection events in personal media collection. An event can be held in the same location more than once but cannot be repeated event in the same time. Therefore once we detect temporal boundaries of an event it became easier to estimate missing spatial information for media entities within the detected event. That makes the event metaphor important for the reconstruction of spatial information for media with missing geographical coordinates. Moreover, the analysis of spatio-temporal information is computationally cheaper in comparison with the analysis of visual features, since time stamps and GPS coordinatenes can efficently be extracted from the EXIF[2] metadata embedded in digital images.

The paper presents a Event-based Semantic Interpolation (EBSI) approach including two steps:

1. Detection of events and their temporal boundaries within an unsorted and not tagged personal media collection.

[1] Global Position System

[2] http://www.exif.org/

.

2. Assigning missing GPS information for each sample within the temporal boundaries of each event. This is performed by interpolation or extrapolation techniques based on temporal distances between samples. For this purpose we use free online navigation services.

Interpolation and extrapolation methods require the presence of geotagged photos within the collection. So we assume that some of the samples in the media collection either were captured by GPS-equipped device (e.g smart phone, camera) or annotated by the owner of the collection.

The rest of the paper organized as follows. Section 2 gives the state of the art, Section 3 presents our approach, Section 4 describes the experimental setup while Section 5 concludes.

2. STATE OF THE ART

Current state of the art techniques for automatic geotagging can be separated on the following categories: visual analysis, text analysis and their combination.

2.1 Visual analysis

Placing an image based only on visual content on global scale is a challenging task. It is difficult to assign location for an image without any context not only for computers but also for humans. At first glance classification of famous landmarks seems solvable to some extend. But considering more generic scenes like sky, forest or indoor images the appropriate geo-annotation become more complex. It happens because of an ambiguity of the image content especially for photos captured indoor. Moreover, *visual analysis* is a significant more time consuming approach than just read the GPS coordinates.

One of the first attempt to place images automatically within the world map is presented in [4]. The proposed approach automatically assigns geo-coordinates for 16% of test images within 200km accuracy. The approach is based on combination of low level features extracted from the training set of geotagged images collected from Flickr[3]. Authors in [5] tackle the problem of placing an image within the urban environment. The work on scene recognition [6] and [7] is related to the image localization task. The work of Hoare et al. [8] presents the approach to triangulate the location of historical images. Their system also able to reconstruct the 3D-model using the old archive photographs.

2.2 Annotation analysis

Any kind of textual description assigned to an image is analyzed in order to estimate its location. In contrast with previously discussed approaches placing images and videos on the map requires user involvement in form of textual description. The process of assigning geographical coordinates to an image based on a given by a user location name is called geocoding. Due to the ambiguity of location names (e.g. Paris, France and Paris, Denmark and Paris Hilton) the problem of distinguishing between them may arise. The problem becomes more complex when a user does not mention any location in the textual description. Authors in [9] approaches the problem of geoannotating by creating language model from user's tags. They place a grid over the world map where each cell on this grid defined by geo-coordinates. The approach is similar to bag-of-word technique. The main idea is to assign set of tags and their scores for each cell in the grid. Laere et al. [10] presents two-step approach where on the first stage they use classifier in order to propose the most likely area where a given photo was captured, and on the next step similarity search is needed to propagate the location with the highest likelihood within the area estimated on the previous step.

2.3 Fusion of textual and visual analysis

The combination of visual and textual modalities recently demonstrated promising results [11]. The framework presented in [12] trains classifier based on combination of textual, visual and temporal features. The authors of the framework point out that photos taken at nearby places and nearly in the same time are probably to be related. It is worth to mention that they limit their task to choose one landmark in the city from a given set of ten examples. [14] presents an hierarchical approach for the task. There, textual and visual modalities are used to determine the region where a video was taken and then - based on visual features - propagated towards geographical coordinates. A similar approach is used in [13]

3. METHODOLOGY

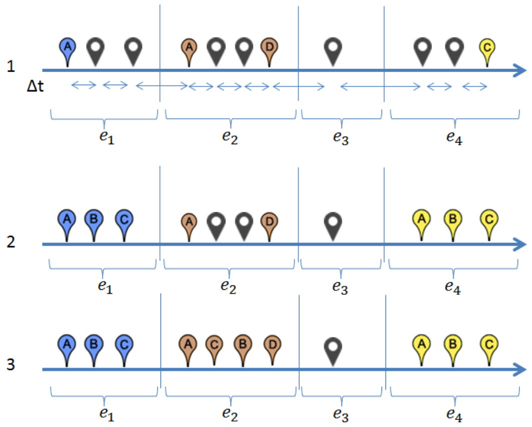

Figure 1: Examples of extrapolation (2) and interpolation (3).

We present an Event-based Semantic Interpolation (EBSI) approach for estimating missing coordinates for images with absent geo information. At the first step the system separates a photo collection on a set of event-related clusters ($e_1 - e_4$) based on temporal information (Δt) only. The detail description of the method for event-based clustering of media presented in [16]. The example is visualized in Figure 1, markers with letters indicate photos with GPS data, dark ones are photos without GPS data. Considering the position of the image in accordance to temporal boundaries there are two possible cases for assigning missing data points:

1. Extrapolation (Figure 1 (2)) is the task of extending a known sequence of values A_{e_1} or C_{e_4} .

2. Interpolation (Figure 1 (3)) is the task of estimation of a unknown sequence of samples within two known data points A_{e_2} and D_{e_2}. The linear interpolation can be described by the formula 1, where the interpolant y can be computed between two point (x_a,y_a) and (x_b,y_b) on a given x.

$$y = y_a + (y_b - y_a)\frac{x - x_a}{x_b - x_a} \tag{1}$$

In the case of extrapolation we extract from the first A_{e_1} or last C_{e_4} geotagged image within an event e_1,e_4 and assign it coordinates to all images without GPS-stamp B_{e_1}, C_{e_1},A_{e_4},B_{e_4} towards the event boundary.

[3] http://www.flickr.com

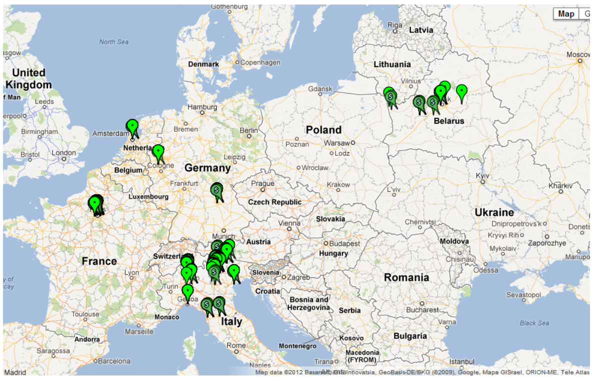

Figure 2: All locations of photos in the data-set automatically assigned by device and assign by EBSI (slanted marker "S").

In case of interpolation we do the following steps. Knowing the coordinates of two points where user made photos (A_{e_2} and D_{e_2}) during the event e_2 EBSI quires on-line navigator in order to understand how user moves between those two points. The are three different variants of travel mode: walking, bicycling and driving. As soon as the travel mode is identified the system queries navigator again. This time it quires the coordinates of a point with the given coordinates of initial point, travel mode and temporal distance to the next sample without coordinates. As the result the semantic analysis is done based on suggestions of travel routes using the Google Maps API[4]. If no route is provided, the locations are linearly interpolated based on temporal distances. In case of absence of geotagged samples within an event e_3 interpolation can be done with help of samples from previous or next event (D_{e_2} and A_{e_4}).

4. EXPERIMENTAL SET-UP

In this section we describe the experimental setup for automatic geotagging of images with missing geographical coordinates. Firstly we will discuss the data set, followed by the experiment description and results.

4.1 Data Set

The data-set consists of 1615 images taken within a period of 1 year and 9,5 months. The data-set was produced unintentionally, meaning the owner was not aware that it would be used for this research. All images have time stamps and 901(55.79%) images have GPS stamps. The images have been captured in six countries and 32 cities and towns. The photos are taken by a Google Nexus One[5] smartphone with a 5MP resolution of 2592 × 1944, sRGB IEC-61966-2 color profile and a fixed focal length of 4,31. For scientific purposes, the data-set is available on request.

The given data-set exemplifies a typical private photo collection. The ground-truth provided by the owner of the collection. The user reconstructed missing spatial information manually with the help of Google Street View[6]. He reported at least 200 m accuracy of placing for each sample. We compared his manual annotation with GPS coordinates automatically assigned to photos by the camera. The results can be seen on the Figure 3. The device is able to place only 71% of images within 200 meters error. The results clearly indicate that GPS reception of the device is not always correct.

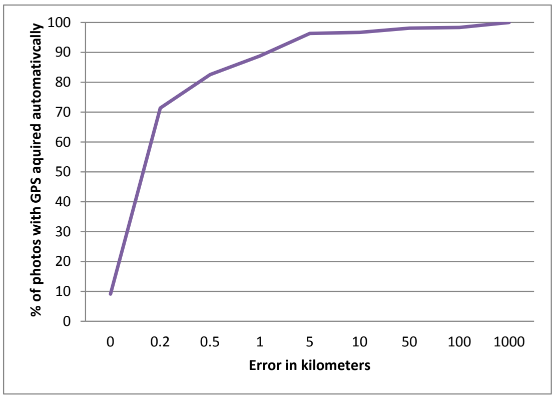

Figure 3: Comparison of images with manual geoannotation and assigned by GPS-enabled device.

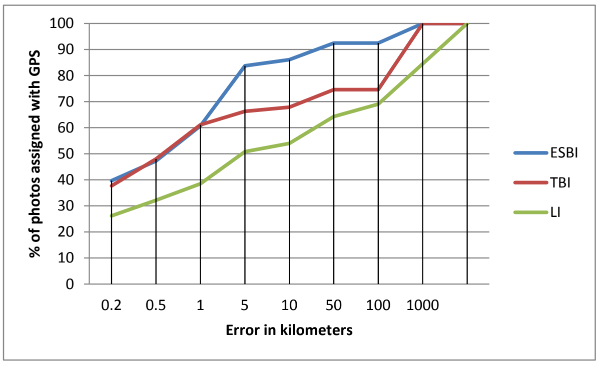

Figure 4: Comparison results for different approaches .

4.2 Experiments

For evaluation of our approach (**EBSI**) we propose to use **linear interpolation** (**LI**) as a baseline. We also tested **temporal based interpolation** (**TBI**) in order to estimate the influence of temporal information for interpolation process. For **TBI** we compute time distances between samples and on their basis perform interpolation. Achieved results presented on the Figure 4 and Table 1. It is worth to mention that EBSI was able to assign geographical coordinates only for 35.5% from the total number of images with missing geo information. This clearly indicates that vast majority of event-related clusters does not even contain a single sample with geo information. For such a case TBI can be used or the user should be involved. TBI and EBSI shows the similar accuracy till 1 km precision and both significantly outperform LI. However from the next threshold EBSI performance increases noticeably. This leap in performance allows to the system automatically place on the global map more than 83% of test images within the 5km error (Figure 2).

[4] https://developers.google.com/maps/documentation/geocoding/
[5] http://www.google.com/phone/detail/nexus-one
[6] http://maps.google.com/help/maps/streetview/

Error in km	0.2	0.5	1	5	10	50	100	1000	>1000
EBSI % of images	39.28	47.22	60.71	83.73	86.11	92.46	92.46	100	100
TBI % of images	37.70	48.02	61.11	66.27	67.86	74.60	74.60	100	100
LI % of images	26.19	32.14	38.49	50.79	53.97	64.29	69.05	84.52	100

Table 1: Experimental results for *Event-Based Semantic Interpolation* (*EBSI*), *Time-Based Interpolation* (*TBI*) and *Linear Interpolation* (*LI*)

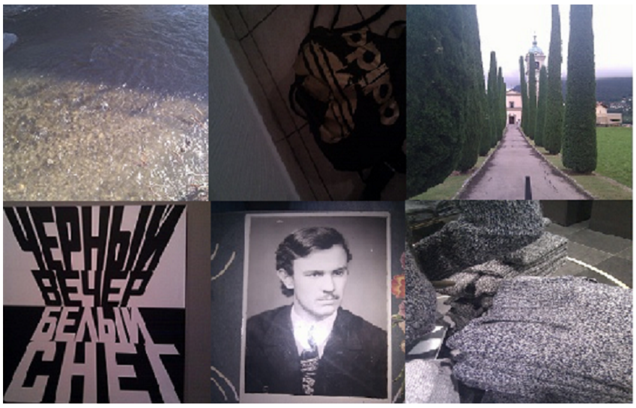

Figure 5: Most accurately interpolated images. EBSI works best if official roads are nearby, since the possible way of traveling is estimated online.

4.3 Conclusion

In this paper we introduce the novel method for automatic geotagging based on the context of personal media collection. Event-based interpolation of images with missing geographical information demonstrates promising results. The approach unveil the significant role of events which they play in reconstruction of missing geo-spatial information. The experiments show that we are able to assign geographical coordinates for 83% of images within an error of 5 km. This is done without looking at the content of the image. In some photos (Figure 5) content information does not provide any cues to distinguish the location where it was captured.

The approach does not require any kind of prior training. However the accuracy of the proposed method highly depends on the number of images with assigned GPS coordinates within the collection. We believe that the combination of contextual, visual and textual information can significantly increase the robustness of the automatic geotagging.

5. REFERENCES

[1] Javier Paniagua, Ivan Tankoyeu, Julian Stöttinger, Fausto Giunchiglia Media Indexing by Personal Events. *ACM International Conference on Multimedia Retrieval (ICMR)*, 2012.

[2] Adam Rae, Vannesa Murdock, Pavel Serdyukov and Pascal Kelm. Working Notes for the Placing Task at MediaEval 2011. *Working Notes Proceedings of the MediaEval 2011 Workshop (MediaEval)*, 2011.

[3] Maarten Clements, Pavel Serdyukov, Arjen P. de Vries, Marcel J. T. Reinders Personalised Travel Recommendation based on Location Co-occurrence. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (IEEE)*, 2011.

[4] James Hays, Alexei A. Efros. IM2GPS: estimating geographic information from a single image. *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[5] Wei Zhang, Jana Kosecka. Image Based Localization in Urban Environments. *Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission. 3DPVT*, 2006.

[6] Aude Oliva , Antonio Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision 3DPVT*, 2006.

[7] Laura Walker Renninger, Jitendra Malikb. When is scene identification just texture recognition? *Vision Research Volume 44, Issue 19,* , 2004.

[8] Cathal Hoare, Humphrey Sorensen. On Automatically Geotagging Archived Images. *Libraries in the Digital Age Proceedings LIDA* , 2012.

[9] Pavel Serdyukov, Vanessa Murdock and Roelof van Zwol. Placing flickr photos on a map. *In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval SIGIR '09* , 2009.

[10] Olivier Van Laere, Steven Schockaert and Bart Dhoedt Finding locations of flickr resources using language models and similarity search. *In Proceedings of the 1st ACM International Conference on Multimedia Retrieval ICMR '11* , 2011.

[11] Martha Larson, Mohammad Soleymani, Pavel Serdyukov, Stevan Rudinac, Christian Wartena, Vanessa Murdock, Gerald Friedland, Roeland Ordelman and Gareth J. F. Jones Automatic tagging and geotagging in video collections and communities. *In Proceedings of the 1st ACM International Conference on Multimedia Retrieval ICMR '11* , 2011.

[12] David J. Crandall, Lars Backstrom, Daniel Huttenlocher and Jon Kleinberg. Mapping the World's Photos. *In Proceedings of the 18th international conference on World wide web WWW'09* , 2009.

[13] Dhiraj Joshi, Andrew Gallagher, Jie Yu and Jiebo Luo Inferring photographic location using geotagged web images. *MULTIMEDIA TOOLS AND APPLICATIONS, Volume 56, Number 1* , 2012.

[14] Pascal Kelm, Sebastian Schmiedeke and Thomas Sikora. A hierarchical, multi-modal approach for placing videos on the map using millions of Flickr photographs. *In Proceedings of the 2011 ACM workshop on Social and behavioural networked media access. SBNMA'11* , 2011.

[15] Pierre Andrews, Jaiver Paniagua and Fausto Giunchiglia. Clues of Personal Events in Online Photo Sharing. *Detection, Representation, and Exploitation of Events in the Semantic Web DeRiVE'11* , 2011.

[16] Ivan Tankoyeu, Javier Paniagua, Julian Stöttinger, Fausto Giunchiglia. Event detection and scene attraction by very simple contextual cues. *Proceedings of the 2011 joint ACM workshop on Modeling and representing events (J-MRE'11)*, 2011.