A Robust Detection of Tourism Area from Geolocated Image Databases

Chareyron Gaël and Da Rugna Jérome

***Abstract*—This paper presents a small part of a project of a framework dedicated to tourism behavior analysis. Several papers explain relation between photographic behavior and touristic behavior. We may find several online image databases allowing users to upload their images and to localize each image on a map, using GPS coordinates. These websites are representative of tourism practices and constitute a proxy to analyze tourism flows. We focus on an automatic method to extract touristic site from large data provided by online image website. We use image signal processing algorithm in our method to extract automatically area of interest. Our paper presents the acquired data and relationship between photographers, sites and photos and introduces the model designed to correctly extract site of interest from data.**

***Index Terms*—Image processing, Segmentation, Tourism flow analysis, mobile application**

I. Introduction

TOURISM for several country is a primordial matter to economy. To help government institution and public tourism travel centers to analyze touristic flow we provide a complete framework. With the wide expansion of smart phone and geolocation technologies, we can use online image website to provide recent data on tourism practice. Indeed, we may find several online image databases allowing users to upload their images and to localize each image on a map. Images and photographies have always representing a tourism and sociological view[1], [2], [3], [4], [5]. Moreover, a recent work [6] has demonstrated that these websites are representative of tourism practices and constitute a proxy to analyze tourism flows. Nevertheless, we have very large amount of data and we need to extract automatically point of interest.

The first part of our paper presents the acquired data and relationship between photographers, sites and photos: one hundred million of geolocated photos using latitude and longitude, each photo owned by one of the 4 million of photographers. The second part shows simple processing and vizualisation. The third part introduces the model designed to automatically extract site and tourist path.

II. Public photo databases are tourism proxy

The increasing use of smartphone, digital camera, Global positioning Systems (GPS), and web-based services in our personal and professional activities are changing the way we communicate and interact with each other but also how we perceive our environment. Now, when we add photos on the web we include geographical informations.

We use two photo-sharing web sites: Flickr and Panoramio. Only the geocoded image data have been recorded on the database. People using these websites have the possibility to add geographical attribute to the image. You can also use devices with integrated GPS to include metadata. Each time a photo is tagged with a physical location, software assigns latitude an longitude values together with an accuracy value derived from the zoom level of the map. Moreover, the system adds metadata embedded by the camera into the image. This information completes the geographical information. All these metadata are saved using Exchangeable Image File Format (EXIF). Table I shows the information available in each photo.

TABLE I
EXIF data recorded in each image.

latitude	longitude	camera model
taken date	shutter speed	aperture value
camera serial number	focal length	make of camera

To extract information we use the public Application Programming Interface (API) provided by Flickr and Panoramio to query the public data store. We choose to download all the data from 2005 until now for all the earth. Each week we crawl new photos to update the database. The table II shows some information on each website.

TABLE II
Data from Panoramio and Flickr overview.

	Panoramio
Photographies Photographer	35 M 1.4 M
Advantages Drawbacks	Only touristic area No complementary information
	Flickr
Photographies Photographer	110 M 900 K
Advantages Drawbacks	Complementary information, ex: place of residence Not only touristic area

In this section, focus is put on the behaviors of tourist in Paris area. Photographers are separated into 2 groups based on their origin: french or not. To find out more about the origin of photographers, we use the information provided by the profile of the user. Many people voluntary provide additional information about themselves such as their city and country of residence. In some case, the origin of the photographer may be also estimated with an high efficiency using a particular algorithm. This algorithm will be described in a future publication. Table III provides information about the photographers in Paris.

III. Simple processing and visualization

To visualize large amount of data from these sources, a geographical representation is adequate[7], [8]. In this context, R, a statistical software, allows to support visual

G. Chareyron and J. Da Rugna are with the Pôle universitaire Léonard de Vinci, ESILV, 12 avenue L. de Vinci, Paris - La Défense, France e-mail: gael.chareyron@devinci.fr.

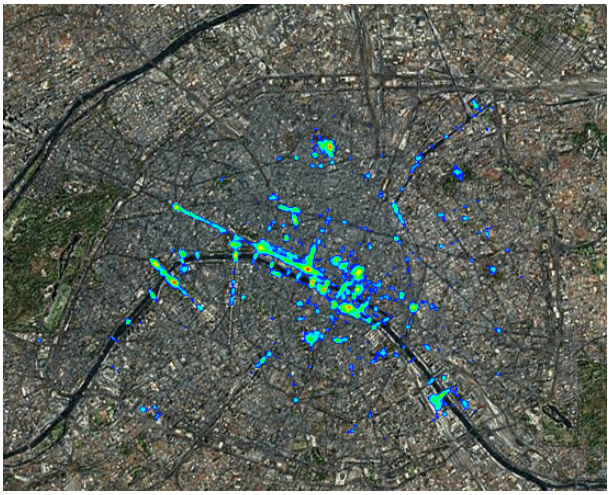
ISBN: 978-988-18210-9-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE III
DATA FOR ILE DE FRANCE (PARIS AREA)

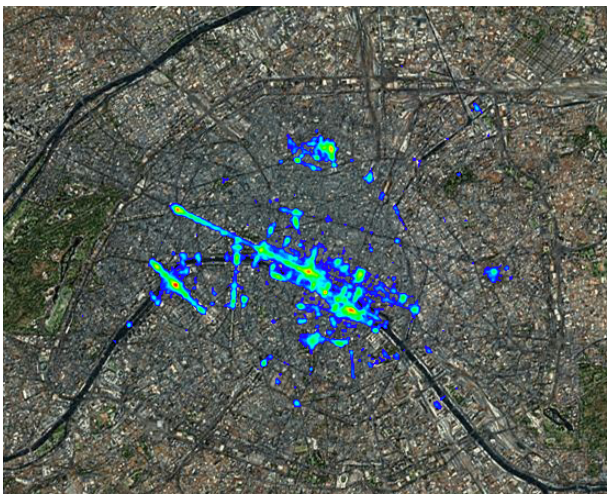
Ile de France		
	Foreigner	French
Photographies	500K	490K
Photographers	17538	5330

synthesis and preliminary investigation of digital traces. All the data are stored in a MySQL© database and R performs visualization and database connection. In addition, another tools were used to create data overlay on satellite maps provided by Google Maps©. To map the spatial distribution of users, data are stored in a matrix covering the entire study of area. Each cell in the matrix includes the number of photographies and unique photographers in this area. Different scales are used, depending on the working area.

Spatial presence can be correlated to the place of residence of the photographer - using Flickr data -. Figure 1 shows the difference between all photographers (a) and foreign photographers(b).



(a)



(b)

Fig. 1. Spatial presence of photographers in Paris. French photographers (a) and foreign photographers (b)

We have seen in this section that internet photography websites could be good proxy of tourism practices. Using this public information, it is possible to help the tourism expert to describe where tourist are, from where they are coming and when they are coming. This is a powerful tool for tourism specialists. Nevertheless, heatmap provide visual information but it's impossible to extract and separate site for a city.

IV. TOURIST SITE DISCOVERY AND REPRESENTATION

We have shown, in the previous section, how public internet sites of geolocalised photographies could be a proxy of tourism practices. This section will introduce a model to extract and manage high-level information from the data. Let us describe the two steps of our model: the site identification and the path identification. In all this section, we consider a working tourism area. This zone is defined by the user, as the zone he wants to visit.

A. Site identification

Figures 2 and 3 illustrate our problematic: how to identify sites in such a data ? Theses figures shows, respectively, the distribution of photographers and photographies on a $0.001° \times 0.001°$grid. More precisely, about 2800 photographs have made at least one photography in a cell near Louvre museum while about 18000 photographies in a cell near Notre Dame of Paris have been made. Photographers near Notre-Dame are concentrated in front of Notre-Dame in a very small area while, Louvre area is large and generates multi-spot, id est point of view for photographies. Thus, exploiting as best as possible this information requires to describe each site by its perimeter and two indicators: representing the photographer distribution, and representing the photography distribution. The site identification is divided in two phases:

- Site localization - To detect where the sites are.
- Site description - To measure the site tourism interest.

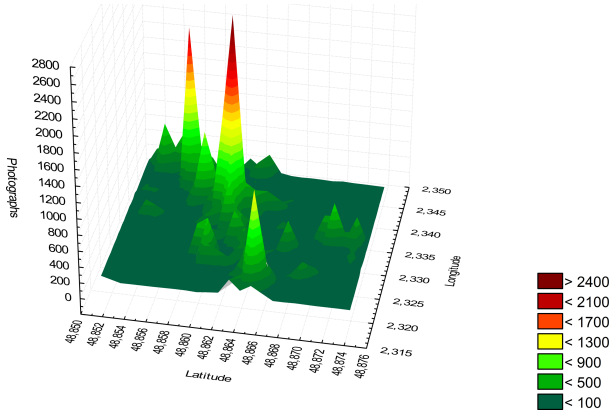


Fig. 2. Photographer number by latitude and longitude in a specific zone in Paris (center near the Louvre museum). The highest peak is the Louvre Museum. Other peaks represent well-known place in Paris: Notre Dame of Paris (the second highest peak), les Tuileries, Place Vendome, place Concorde, etc. Grid discretization step is $0.001°$ for latitude and for longitude.

1) Site localization: As explained above, the two distributions are present and will be used. Let's name UD the photographer distribution and PD the photography distribution. These distributions are obtained using a 2d grid of size 2000×2000. The grid resolution depends only on working tourism area. Considering a small city, it may represent a

ISBN: 978-988-18210-9-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

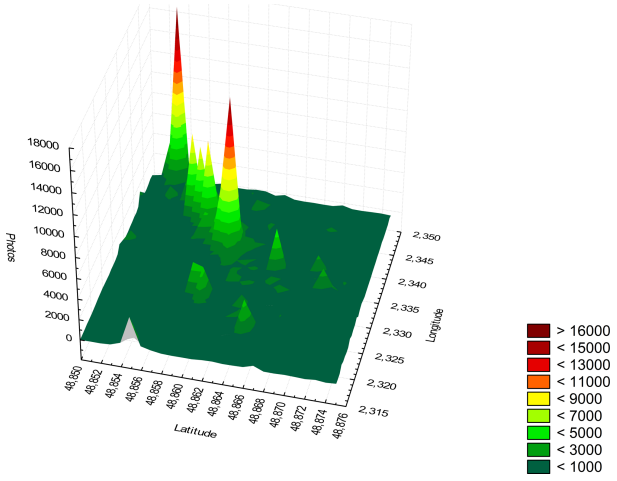


Fig. 3. Photography number by latitude and longitude in a specific zone in Paris (center near the Louvre museum). The highest peak is the Notre Dame of Paris, the second Louvre museum. Grid discretization step is 0.001° for latitude and for longitude.

step of 0.001°, considering a state, it may represent a step of 0.01°. Thus, these distributions represent 4 millions of potential sites. Detecting area in a 2D data may be considered as an image processing process on a grey level images. In this context, the peak-finding algorithm 1 introduced by Cheng et al[9] is used to identify the most significant peaks in each distribution. α is a threshold used to exclude not enough representative peaks and β represents the minimum distance allowed between two peaks[1]. To avoid high-frequency effect, a low-pass filter is applied using a Gaussian kernel.

Algorithm 1: Peak-finding algorithm.
Input: A normalized grid distribution H
Output: Significant peaks of the distribution
$H\prime \leftarrow H \star$ (Gaussian Kernel)
$Peaks \leftarrow$ Local maxima of $H\prime$
$Peaks \leftarrow$ Local maxima of $Peaks$
$T_\alpha \leftarrow \alpha.max(Peaks)$
$Peaks \leftarrow \{p \in Peaks; H\prime(p) \geq T_\alpha\}$
foreach $(p_1, p_2) \in Peaks \times Peaks$
  if $\|p_1, p_2\| \leq \beta$
    if $H\prime(p_1) < H\prime(p_2)$
      $Peaks \leftarrow Peaks \backslash \{p_1\}$
    else
      $Peaks \leftarrow Peaks \backslash \{p_2\}$

Considering normalized (from 0 to 1) distributions UD and PD, this algorithm is applied to 5 distributions: 5 peak sets are extracted.

- UD and PD
- $0.25 \times UD + 0.75 \times PD$ and $0.75 \times UD + 0.25 \times PD$
- $0.5 \times UD + 0.5 \times PD$

Tourism sites defined are peaks present in, at least, 2 peak sets. Figure 4 exhibits detected site centers in Paris center.

2) Site description: To describe a site, it requires to define the influence area of each site. Figure 5 shows, in 1 dimension, different types of peaks: small, large, high or not...To achieve an efficient peak description, a local

[1]Thresholds are relative to area zone size

(a) Area of interest for french

(b) Area of interest for french

Fig. 4. Example of the area-finding algorithm in center of Paris.

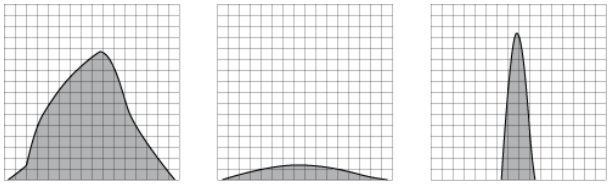


Fig. 5. 1D illustration of peak diversity and importance to compute site influence area.

watershed algorithm[10], [11] is used and allow to obtain, for each peak, an attached perimeter. Still considering UD and PD, $UD_{x,y}$ and $PD_{x,y}$ are defined as the distribution value in sub-square at position (x,y) in the 2000×2000 grid. A site S is then defined by all its sub-squares. Let's define $\gamma(S), \Gamma, \delta(S)$ and Δ:

$$\gamma(S) = \sum_{(x,y) \in S} UD_{x,y} \quad (1)$$

$$\Gamma = \sum_{(x,y)} UD_{x,y} \quad (2)$$

$$\delta(S) = \sum_{(x,y) \in S} PD_{x,y} \quad (3)$$

$$\Delta = \sum_{(x,y)} PD_{x,y} \quad (4)$$

ISBN: 978-988-18210-9-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Thus, two interest measures are computed for each site.

$$I_{UD}(S) = \frac{\gamma(S)}{\Gamma} \quad (5)$$

$$I_{PD}(S) = \frac{\delta(S)}{\Delta} \quad (6)$$

$I_{UD}(S)$ measure the probability to visit a site and $I_{PD}(S)$ the probability to take a photography in this site.

B. Path identification

Intelligent tourism application proposes to the tourist the ability to select the best site to visit. To find it, this information is present:

- user path - This is the GPS trace of the user. This path is reduced to site path: every sites where the user was. Let's call this trace $T = \{T_j,\ 1 \leq j \leq N\}$. T_j is the j-th site visited by the user.
- site description - S, $I_{PD}(S)$ and $I_{UD}(S)$ for each site S. Let's call M the number of sites and, $S = \{S_k,\ 1 \leq k \leq M\}$ the set of sites.
- internet user tourism path - For each user of database, the path in the visiting area is considered. Let's call these traces: $\Theta = \{\Theta^l_j, 1 \leq l \leq R, 1 \leq j \leq Q_l, \}$. R is the number of visitors in visiting area, Q_l the number of visited sites by the l-th visitor. Θ^l_j the j-th sites of the l-th visitor trace.

1) Simple selection: To find the next site to visit, it requires to compute, for each unvisited site, a potential interest estimator $E(S_k)$. A first and naive approach is to directly use I_{PD} or I_{UD}:

$$E(S_k) = \begin{cases} 0 & if\ S_k \in T \\ \Phi(I_{PD}(S_k), I_{UD}(S_k)) & else \end{cases} \quad (7)$$

where $\Phi(I_{PD}(S_k), I_{UD}(S_k))$ is a combination of $I_{PD}(S_k)$ and $I_{UD}(S_k)$, like shown in table IV.

TABLE IV
SEVERAL POSSIBLE DEFINITIONS OF $\Phi(I_{PD}(S_k), I_{UD}(S_k))$. BY SELECTING A CHOICE, THE USER DEFINES A PRIORITY OR NOT BETWEEN PHOTOGRAPHY DENSITY AND PHOTOGRAPHER DENSITY.

$I_{PD}(S_k)$	$I_{UD}(S_k)$
$0.5 \times I_{PD}(S_k) + 0.5 \times I_{UD}(S_k)$	$MAX(I_{PD}(S_k), I_{UD}(S_k))$

Sites maximizing $E(S_k)$ are then selected and shown to the user as potential interesting sites. $E(S_k)$ also indicates a measure of interest useful for the user.

2) Advanced selection: The previous measure of potential interest is very simple but only efficient considering short trace: if user have visited only 2 or 3 sites, the best choice may be to select most visited sites. Nevertheless, a longer trace have to be exploited: $E(S_k)$ must integrate the information Θ. Let us first define a measure of weighted distance between user trace and database trace. Complex, more adapted, distances or algorithms, like Earth mother distance, were not used to keep computation time minimal[12], [13].

$$D_l = \sum_{S \in T \setminus \{\Theta^l \cap T\}} \lambda + (1-\lambda) \times \Phi(I_{PD}(S), I_{UD}(S))$$

λ is a weight set between 0 and 1. $\lambda = 0$ means that D_l is the number of sites not present in the two trace intersection. $\lambda = 1$ means that D_l gives priority to high density sites. $\lambda = 1$ will select the main stream sites while $\lambda = 0$ will also manage untypical tourism path. Hence, the nearest trace set Ω is retained: it contains one or more traces, each at the same minimal distance. Finally, $E(S_k)$ is computed as follow:

$$E(S_k) = \begin{cases} 0 & if\ i \in \{T \cup \Omega\} \\ \Phi(I_{PD}(S_k), I_{UD}(S_k)) & else \end{cases}$$

In this section, we have defined how sites are localized and represented. We have also presented a process to search the most interesting sites to visit. Let us present in some words the mobile application.

V. MOBILE DEVICE APPLICATION

To illustrate our study, a mobile device application is proposed for Android systems. Let's describe the primary process, as illustrated figure 6.

1) User runs the application and defines its visiting area: a city, a county, a state, ...
2) Application, over internet, connects to the server and creates a new profile.
3) The server computes peak description in background.
4) The application runs in background and produces the user trace.
5) After a while, user asks for the next site: the trace is sent to the server. The server computes $E(S_k)$.
6) Potential sites are shown. Sites are named using wikipedia reverse geocoding tool.

Fig. 6. Main steps of the mobile application: (a) Along the user visits, application build the gps trace. (b) the trace is sent to the server that computes distance to select sites. (c) Results are shown to the user.

VI. CONCLUSION AND FUTURES WORKS

This work presents a new way to detect touristic sites in a place. Using dedicated image processing tools, we are able to identify the most visited sites and, also, less visited but significant sites. We illustrate our method by developping a mobile application designed to geolocated devices. This application permits the travel guide to match the user wishes. At each moment of the travel of a tourist, the guide adapts the site interest by the tourist history. Also, the tourist is able to define his preferred behavior: main stream or atypical. Furthermore, this guide is never out-dated because data are constantly updated from several sources. This application shows how connected mobile devices and public information may be correlated to produce new and innovative service to the user.

ISBN: 978-988-18210-9-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

The semantic description of each tourism trace is one of you future works. The visited site list is enough for a basic tourism guide, but quality issues will be resolved by integrating other information like: temporal path, season, symbolic description of sites,... This program is a part of a research project on emerging tourism flows, in particular the Paris 2030 program.

REFERENCES

[1] R. M. Chalfen, "Photograph's role in tourism : Some unexplored relationships," *Annals of Tourism Research*, vol. 6, no. 4, pp. 435–447, 1979. [Online]. Available: http://www.sciencedirect.com/science/article/B6V7Y-46BHYJD-JX/2/e0b3f137339874655e54be4722c8457f

[2] M. F. Goodchild, "Citizens as sensors: the world of volunteered geography," *GeoJournal*, vol. 69, no. 4, pp. 211–221, 2007. [Online]. Available: http://www.springerlink.com/content/h013jk125081j628/

[3] K. Grunfeld, "Integrating spatio-temporal information in environmental monitoring data–a visualization approach applied to moss data," *The Science of the Total Environment*, vol. 347, no. 1-3, pp. 1–20, Jul. 2005, PMID: 16084963.

[4] E. O'Neill, V. Kostakos, T. Kindberg, A. F. gen. Schieck, A. Penn, D. S. Fraser, and T. Jones, "Instrumenting the city: Developing methods for observing and understanding the digital cityscape," in *Ubicomp*, ser. Lecture Notes in Computer Science, P. Dourish and A. Friday, Eds., vol. 4206. Springer, 2006, pp. 315–332.

[5] N. B. Salazar, "Imaged or imagined? cultural representations and the tourismification of peoples and places," pp. 49–72, 2009.

[6] G. Chareyron, S. Cousin, J. Da-Rugna, and D. Gabay, "Touriscope: map the world using geolocated photographies," in *IGU meeting, Geography of Tourism, Leisure and Global Change*, 2009.

[7] J. Wood, J. Dykes, A. Slingsby, and K. Clarke, "Interactive visual exploration of a large spatio-temporal dataset: reflections on a geovisualization mashup," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1176–1183, Dec. 2007, PMID: 17968062.

[8] P. Dourish and A. Friday, Eds., *UbiComp 2006: Ubiquitous Computing, 8th International Conference, UbiComp 2006, Orange County, CA, USA, September 17-21, 2006*, ser. Lecture Notes in Computer Science, vol. 4206. Springer, 2006.

[9] H. Cheng and Y. Sun, "A hierarchical approach to color image segmentation using homogeneity," *IEEE Transactions on Image Processing*, vol. 9, no. 12, pp. 2071–2082, December 2000.

[10] L. Vincent and P. Soille, "Watersheds in digital spaces: an efficient algorithm based on immersion simulations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 583–598, 1991.

[11] J.-B. Kim and H.-J. Kim, "Multiresolution-based watersheds for efficient image segmentation," *Pattern Recognition Letters*, vol. 24, pp. 473–488, 2003.

[12] R. Duda, P. Hart, and D. Stork, *Pattern Classification (Second Edition)*. Wiley-Interscience, 2001.

[13] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

ISBN: 978-988-18210-9-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)