What you mark is where it was.
A user-centric approach for post-hoc georeferencing of urban terrestrial photos

Holger Fritze
University of Münster
Institute for Geoinformatics
Münster, Germany
h.fritze@uni-muenster.de

Auriol Degbelo
University of Münster
Institute for Geoinformatics
Münster, Germany
degbelo@uni-muenster.de

Steffen Loos
University of Münster
Institute for Geoinformatics
Münster, Germany
s.loos@uni-muenster.de

Abstract

Geotagging user-generated photos is in many cases not sufficient to properly describe their context. Besides not telling in which direction the photo is facing or what is seen in the photo, estimating the actual location during post-hoc georeferencing is cumbersome and imprecise. This paper presents an approach for georeferencing urban photos by marking at least three distinctive features in the photo and connecting them to their real-world locations using a map. Based on these connections, an algorithm calculates the camera viewshed, its orientation, and location. Algorithmic tests under ideal conditions showed high accuracy rates on a centimeter level with average run-times of ten minutes, as well as accuracy rates of 1.5 meters with average run-times of about one second. We developed a web application which helps users to create the connections using two different input types: points and lines. The user study showed accuracy rates of 9 meters in average. The approach is a user-friendly way of generating useful metadata and is relevant to address the issue of metadata scarcity in online photo sharing platforms.

Keywords: georeferenced media, post-hoc georeferencing, metadata, user-generated content

1 Introduction

User-generated photographs play an increasingly important role when conveying information in online communications. Querying them according to their geographic location require the presence of useful location-related metadata. However, these metadata are not always available. For instance, Li et al., (2013) mention that the percentage of geotagged photos in Flickr is about four percent. There are at least two strategies to improve this situation: develop methods to generate location-related metadata when the photographs are taken (i.e., geotag photos automatically), or elaborate techniques which help to infer location-related metadata *a posteriori* (i.e., geotag photos manually). The focus of the paper is on the latter strategy, and the goal is to provide a semi-manual method to facilitate post-hoc georeferencing of photographs.

Tags generated by users to describe photographs can be classified into three types: non-visual metadata, perceptual descriptions, and conceptual descriptions (Hollink et al., 2004). Non-visual metadata provide descriptive information about the context of the generation of the photographs (e.g., creation date, location, and title). The latter two types relate to the content of the photo. Perceptual descriptions refer to descriptions directly derived from the visual characteristics of the photographs (e.g., color, shape, and texture); conceptual descriptions give information about the semantic content of the photographs (e.g., objects and scenes portrayed).

Previous studies suggest that people rarely provide nonvisual metadata while tagging photographs and videos. Hollink et al., (2004) found only 0.9 % of the user tags to be nonvisual, and Gligorov et al., (2011) found no tag at the non-visual level in their study. This indicates the need for approaches to infer non-visual metadata from user inputs.

This paper introduces an approach to generate camera position, and camera orientation a posteriori. The approach is based on users specifying corresponding features in the photographs and in a map. A preliminary evaluation with seven participants showed that the approach produces accurate results (i.e., 9 m) within a reasonable amount of time (i.e., about one minute per photograph).

2 Related Work

Georeferencing photos traditionally relate to the domain of remote sensing. For orthographic photos captured from aerial platforms, it is a well-established approach to connect and register the pictorial information with corresponding features. This process of georectification assigns each pixel value to a coordinate in a given reference system and applies distortions to adjust to topographic variations of the earth surface.

Since the rise of user-generated content, the term georeferencing is used with different meanings. For terrestrial imagery such as user-generated photos, the term refers to assigning a location information, i.e., a single coordinate pair, to the photo as a whole, and the term is often used interchangeably with "geotagging" and "geolocating" (Hill, 2009). Many social media platforms utilize this basic spatial information of the photos. Panoramio and Flickr are prominent examples of photo sharing platforms utilizing the geotag (Zielstra & Hochmair, 2013). The geotag is often automatically created by built-in functionalities of modern devices and can be derived from e.g., GPS, cell tower IDs, or IP addresses (Hahmann et al., 2015). However, prior research has shown that actually the minority of

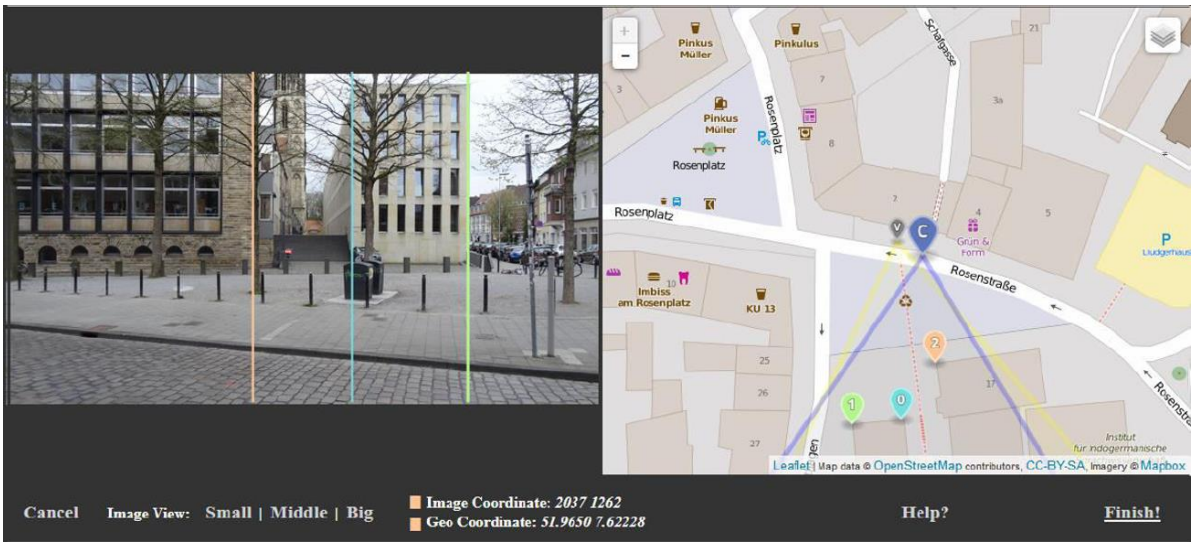

Figure 1: Point method. The photo features F(x,y) used for georeferencing are visualized by vertical lines. The corresponding color markers in the map (0, 1, and 2) refer to the marked geolocations by the user. The marker C illustrates the computed geolocation based on (0,1,2). The smaller marker V shows the actual geolocation, measured by an accurate GPS device, as validation information.

shared photos is geotagged. While Panoramio requires a geotag to upload photos, in Flickr only about four percent of all photos have a geolocation (Alivand & Hochmair, 2016; Li et al., 2013). If a photo has not been annotated with location information at the time it was taken, the associated meta-information (e.g. title, description, and keyword tags) can be used to estimate a location on a post-hoc basis (Estima & Painho, 2013; Sester et al., 2014; Van Laere et al., 2010, 2013). Yet, the inferred location is often not correct. Shared photos on Flickr show geolocation displacements from their actual location of 11 to 13 meters for popular venues and displacements of 47 to 167 meters for unpopular venues (Hauff, 2013). This not only impedes querying photos based on their spatial properties but also complicates the orientational alignment of the photo to the map content and may result in misperceptions. Furthermore, the geotag usually does not provide further information on e.g., the altitude, the orientation, or any non-visual metadata.

Google Street View[1], Mapillary[2], and OpenStreetCam[3] are three examples of applications using the orientation information of terrestrial photos taken at street level. All three applications provide transitions and navigation between the photos and connect them to a map interface providing a spatial and contextual embedding. For the crowd-sourced imagery, Mapillary uses computer vision approaches to correct spatial information and displacements of photos and to align them with the street network (Juhász & Hochmair, 2016).

3 Approach

Our approach of georeferencing urban terrestrial photos goes beyond the geotag and estimates the location of capturing the photo by a post-hoc triangulation of visible features in the photo and their corresponding actual geolocations. We developed an algorithm which infers the camera position and orientation from these manually created correspondences.

For reasons of initial simplifications, we restricted the approach to photos of prominent buildings captured in landscape orientation, parallel to the surface and only rotated around the y-axis. The algorithm requires at least three correspondences consisting of two reference points each: one in the photo and one in a map. The user creates these correspondences through a web application and two different methods. In the point method (cf. Figure 1), she creates a reference point by clicking

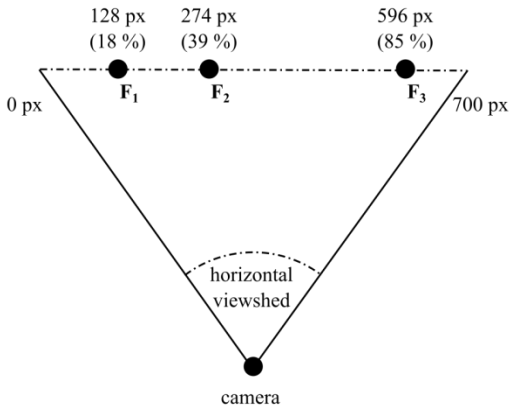

Figure 2: Sample triangulation of comparing the pixel positions of three marked features (F_1 – F_3) from the current camera position. The given image width is 700 px.

[1] www.google.com/streetview (vis. 05.04.2017)
[2] www.mapillary.com (vis. 05.04.2017)
[3] www.openstreetcam.com (vis. 05.04.2017)

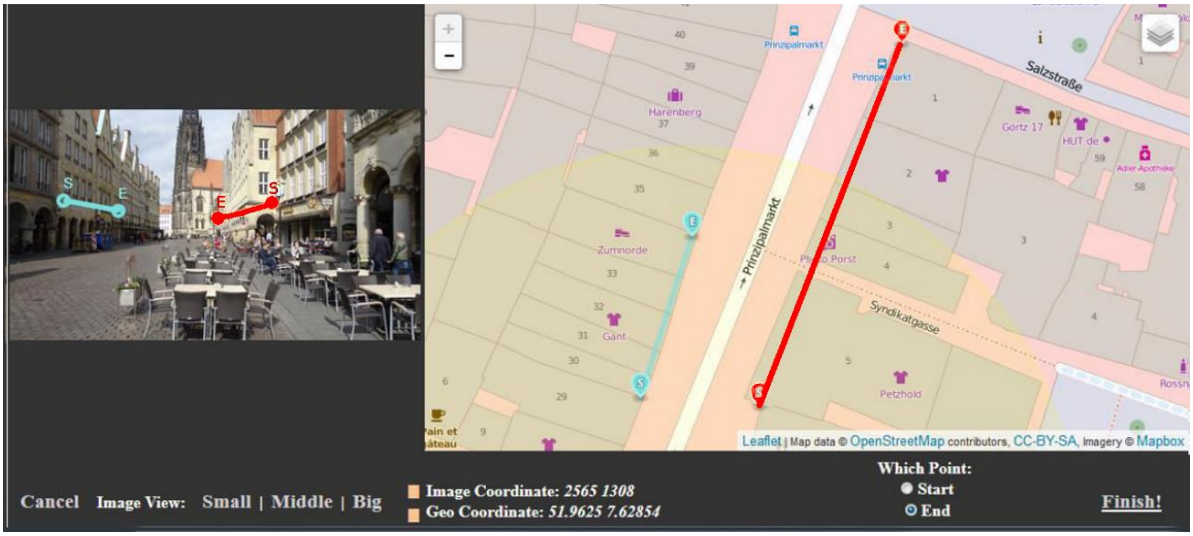

Figure 3: Line method. The photo features F(x,y) used for georeferencing are visualized by horizontal lines along building facades. The markers in the map refer to the (marked) geolocations of the start and end point of the lines.

in the photo and then in the map. The reference point is displayed by a vertical line in the photo and by a marker on the map. Due to the restriction regarding the rotation, we only work with the x-coordinate of the photo, which is set in relation to the photo width. For the second method, the line method, the user also creates points on both sides (cf. Figure 3). But here, the points are visualized as start and end points for lines. For this method, the user is intended to mark along building facades.

The algorithm uses the photo metadata (i.e., the image width and the horizontal viewshed angle derived from the focal length) and the correspondences to compute the set of camera transformations. It then searches for the transformation, where the calculated camera position has the closest match to the marked x-coordinates in the photo. Figure 2 shows the triangulation of the algorithm to calculate where the geolocations of the features refer to their marked positions in the photo. The outer lines illustrate the horizontal viewshed, which is calculated by the focal length of the camera and the sensor width.

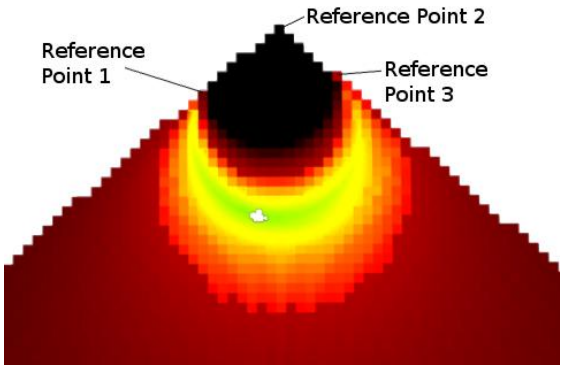

Figure 4: Optimization steps: Only positions within the colored area will be considered by the algorithm if reference points 1-3 are provided as input. The step sizes for the position of higher deviation (red) are much bigger for those of lower deviation (green).

The correct camera's position is estimated numerically by computing the potential camera location, and matching it to the image x-coordinates of the referenced features in the photo. Therefore, we have to consider each possible geolocation with each possible rotation within a certain distance to the features. The algorithm adopts the transformation with the lowest deviation between marked and calculated x-coordinates of the features in the photo (cf. Figure 4). The algorithm is provided as pseudocode and as full Node.js implementation on Github[4].

4 Evaluation

To test the approach and the algorithm, we conducted two evaluations. A technical evaluation tested the runtime and accuracy of the algorithm with and without the optimizations (i.e., skip logically impossible areas, define minimum and maximum distances). A user study compared the point method and the line method as two means of creating the correspondences.

Technical Evaluation

We built a test environment with the game engine Unity3D. In this environment, the marked x-coordinates in the photos fit exactly their actual geolocations. We tested the transformation finding algorithm with two different optimization settings. Without the optimizations, the algorithm showed an average accuracy of about 0.03 m for the location error and 0.03° for

[4] https://github.com/sitcomlab/What-you-mark-is-where-it-was

the orientation error, but needs about 10 minutes computation time. We applied optimizations to skip unnecessary calculations. First, we excluded areas, where a valid camera transformation is logically not possible (cf. Figure 4, non-colored area) and defined a minimum and maximum distance to the buildings. Second, we applied an iterative decrease of the step size. The algorithm iterates through each position with an initial step size of, e.g., 10 meters. For the positions with the lowest deviation in the transformation, the step size is incrementally decreased (cf. Figure 4). The optimizations reduce the average accuracy to about 0.7 m for the location and 0.4° for the orientation, but also reduce the computation time of the algorithm to about one second.

User Study

While the test environment can rely on reference points free of errors, the semi-manual approach relies on user input, which come naturally with marking errors. As the precise creation of the reference points has an important influence on the accuracy of the approach, we compared the point method and the line method during a user study.

Study design: the objective of the study is to get initial insights about the performance and the applicability of both methods for the post-hoc georeferencing of photos. Seven participants were asked to georeference seven photos by applying the point and the line method in the web application introduced in Section 3. All seven photos show familiar points of interest in Münster, Germany. Since georeferencing requires the ability to establish correspondences between features of a photograph and features of a map, the participants were limited to residents of Münster. The study had three steps: (i) a self-assessment of the participants' familiarity with online mapping services in a short questionnaire; (ii) reading of instructions to familiarize themselves with the two methods of georeferencing; and (iii) the georeferencing of the seven photos. A pilot study with two participants was conducted to test the experimental design. The results of this pilot study are not included in the results below.

Validation Data: to assess the algorithm's accuracy, we collected ground truth information on the camera's position and orientation using a highly accurate GPS device (Garmin GPSmap 60SCx).

Table 1: Results of the user study (with values rounded to two decimal places). The point method is faster, but yields more errors regarding the camera position and orientation. The georeferencing time describes the duration of interaction, and includes the creation of the correspondences and the processing of the algorithm.

Method	Point	Line	Average
georeferencing time	53 s	63 s	58 s
camera position error	10.61 m	7.56 m	9.12 m
camera orientation error	14.06°	9.16°	11.66°

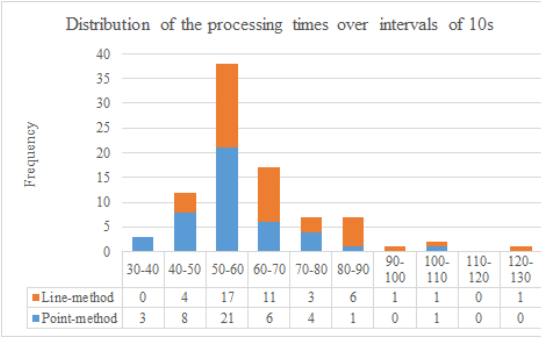

Figure 5: Distribution of the georeferencing times.

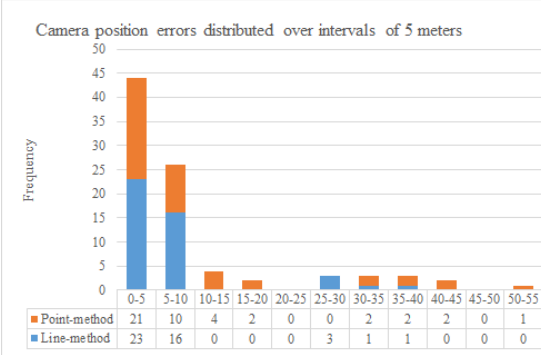

Figure 6: Distribution of the camera position errors. The data from the point method is much more dispersed than the data from the line method.

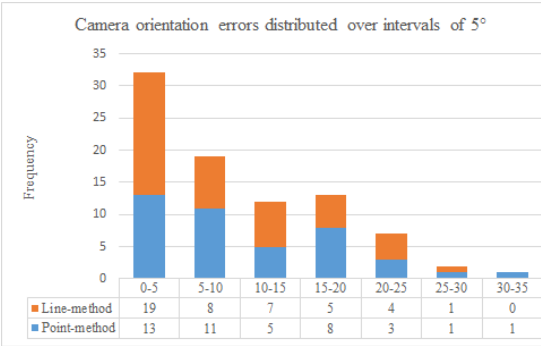

Figure 7: Distribution of the camera orientation errors. Here also, the data from the point method is much more dispersed than the data from the line method.

Results: the seven study participants were between 19 and 27 years old. All confirmed to know at least four of the six points of interest presented in the seven photos. In total 49 values of camera position and camera orientation were collected during the study. Table 1 shows the average time needed to apply each method, as well as the location and orientation errors.

Figure 5 shows the data points of the processing time spread over intervals of 10 s. The standard deviation is 16.3 s for the point method and 20.1 s for the line method. Figure 6 shows data points of camera position errors spread over intervals of five meters each. Here, the standard deviation is 13m for the point-method and 8.8m for the line method. Figure 7 shows data points of the camera orientation errors, spread over intervals of 5°. The standard deviation is 18.2° for the point method and 7.3° for the line method.

5 Conclusion and Outlook

This article introduced a user-centric approach for the post-hoc georeferencing of urban photos. We developed an algorithm to compute the geolocation and orientation of the camera at the moment of capturing the photo, based on user inputs. A user creates correspondences between the photo and the real-world. The algorithm can compute highly precise results at the cost of a long processing time (positional accuracy of 3 cm with approx. 10 min run time). Optimizations could reduce the processing time and still preserve the accuracies at a feasible level (positional accuracy of 1.5 m with approx. 1 s run time). The user input has been tested applying two methods for creating the correspondences between the photo and a map and showed average accuracies of nine meters (average georeferencing time: 58 s). The reasonable amount of time indicates that both the point and the line method are worth considering during the post-hoc georeferencing of photographs. Furthermore, the data suggest that the point method is (a) overall faster, (b) more stable with respect to processing time; but (c) less homogeneous (and therefore predictable) as to errors.

These results are promising but need further investigations in follow-up studies before definitive statements can be made regarding the respective merits of the two methods. The user study also showed the positional and orientational errors to be primarily caused by inaccurately placed reference points on the photo as well as on the map. This underlines the need for high image resolutions to more precisely place the reference points, as well as adequately designed user interfaces to support a user-friendly post-hoc georeferencing of existing urban photos.

Acknowledgments

Auriol Degbelo gratefully acknowledges funding from the European Union through GEO-C (H2020-MSCA-ITN-2014, 642332, http://www.geo-c.eu/).

References

Alivand, M., & Hochmair, H. H. (2016). Spatiotemporal analysis of photo contribution patterns to Panoramio and Flickr. *CaGIS*, *44*, 170–184.

Estima, J., & Painho, M. (2013). Flickr Geotagged and Publicly Available Photos: Preliminary Study of Its Adequacy for Helping Quality Control of Corine Land Cover. In B. Murgante et. al (Eds.), *Computational Science and Its Applications – ICCSA 2013* (pp. 205–220).

Gligorov, R., Hildebrand, M., van Ossenbruggen, J., Schreiber, G., & Aroyo, L. (2011). On the Role of User-generated Metadata in Audio Visual Collections. In *Proc. 6th Int. Conf. Knowledge Capture* (pp. 145–152).

Hahmann, S., Purves, R., & Burghardt, D. (2015). Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. *Journal of Spatial Information Science*, *0*, 1–36.

Hauff, C. (2013). A Study on the Accuracy of Flickr's Geotag Data. In *Proc. 36th Int. ACM SIGIR Conf. on R&D in Information Retrieval* (pp. 1037–1040).

Hill, L. L. (2009). *Georeferencing: the geographic associations of information*. Cambridge, Mass.: MIT Press.

Hollink, L., Schreiber, A. T., Wielinga, B. J., & Worring, M. (2004). Classification of user image descriptions. *International Journal of Human-Computer Studies*, *61*, 601–626.

Juhász, L., & Hochmair, H. H. (2016). User Contribution Patterns and Completeness Evaluation of Mapillary, a Crowdsourced Street Level Photo Service: Contribution Patterns of Mapillary. *Transactions in GIS*, *20*, 925–947.

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *CaGIS*, *40*, 61–77.

Sester, M., Arsanjani, J. J., Klammer, R., Burghardt, D., & Haunert, J.-H. (2014). Integrating and Generalising Volunteered Geographic Information. In D. Burghardt et. al. (Eds.), *Abstracting Geographic Information in a Data Rich World* (pp. 119–155). Springer.

Van Laere, O., Schockaert, S., & Dhoedt, B. (2010). Towards Automated Georeferencing of Flickr Photos. In *Proc. 6th Workshop on Geographic Information Retrieval* (pp. 5:1–5:7).

Van Laere, O., Schockaert, S., & Dhoedt, B. (2013). Georeferencing Flickr resources based on textual meta-data. *Information Sciences*, *238*, 52–74.

Zielstra, D., & Hochmair, H. H. (2013). Positional accuracy analysis of Flickr and Panoramio images for selected world regions. *Journal of Spatial Science*, *58*, 251–273.