Estimating Poses of World's Photos with Geographic Metadata

Zhiping Luo, Haojie Li, Jinhui Tang, Richang Hong, and Tat-Seng Chua

School of Computing, National University of Singapore
{luozhipi,lihj,tangjh,hongrc,chuats}@comp.nus.edu.sg

Abstract. Users can explore the world by viewing place related photos on Google Maps. One possible way is to take the nearby photos for viewing. However, for a given geo-location, many photos with view directions not pointing to the desired regions are returned by that world map. To address this problem, prior know the poses in terms of position and view direction of photos is a feasible solution. We can let the system return only nearby photos with view direction pointing to the target place, to facilitate the exploration of the place for users. Photo's view direction can be easily obtained if the extrinsic parameters of its corresponding camera are well estimated. Unfortunately, directly employing conventional methods for that is unfeasible since photos fallen into a range of certain radius centered at a place are observed be largely diverse in both content and view. Int this paper, we present a novel method to estimate the view directions of world's photos well. Then further obtain the pose referenced on Google Maps using the geographic Metadata of photos. The key point of our method is first generating a set of subsets when facing a large number of photos nearby a place, then reconstructing the scenes expressed by those subsets using normalized 8-point algorithm. We embed a search based strategy with scene alignment to product those subsets. We evaluate our method by user study on an online application developed by us, and the results show the effectiveness of our method.

1 Introduction

Google Maps is a widely used online service to explore world's places. However, current service mainly relies on the geographical metadata of photos result in simply considering the photos nearby a place are exactly related to the geographic content. We can observe the limitations of such application. On one hand, photos taken by non location-aware devices may be wrongly placed on the map by uploaders manually. On the other hand, even the photos are correctly placed (manually by users or automatically from EXIF tags), their viewing directions may not be pointing to the desired region. Therefore, many photos without truly poses in terms of position and view direction are returned.

Prior know the poses in terms of position and view direction of photos, then let Google Maps returns only nearby photos with view direction pointing to the target places is one of the most feasible way to address the above problems.

S. Boll et al. (Eds.): MMM 2010, LNCS 5916, pp. 695–700, 2010.
© Springer-Verlag Berlin Heidelberg 2010

Assume the geographic metadata is correct, It can be easy to get the photo's position expressed by latitude and longitude coordinate referenced by a world map from the metadata. In this paper we consider Google Maps be the case as its popularity of use. We then can further obtain the photo's view on the map by geo-registering its view direction estimated in the camera coordinate system. To estimate the view direction, we can first estimate the camera rotation and translation via scene reconstruction. When there are significant overlaps among photos [1], rotation and translation can be robustly estimated.

The most used technique for scene reconstruction is described as follows. First compute correspondences among images based on feature matching. Second, employ RANSAC [2] to decide the inliers (the actual correspondences) when tuning the estimation of fundamental matrix [3]. At last, compute the camera extrinsic parameters, such as rotation and translation, based on the assumption of fixed intrinsic parameters.

However, those photos in the Internet that we can easily obtain are largely diverse both in image content and view. We randomly selected 10 photos fallen into the range of 20-meter radius centered at a region within "Acropolis, Athens, Greece", and calculate the inliers among them. unfortunately, Only average 40% inlier rate are obtained by point-to-point matching SIFT [4] features with well configure and under 10^6 iterations in RANSAC for tuning the estimations of fundamental matrices. Since current techniques for scene reconstruction heavily depend on inlier rate so that such a low value cannot satisfy the estimation of camera's extrinsic parameters well. Since scene reconstruction can be well done if there are significant overlaps among photos, We propose to divide the whole photos a number of subsets. In each set, the photos are more convinced to be visually relevant, meanwhile they represent an underlying scene consisting of those photos. Therefore, those underlying scenes might be well reconstructed since their own photos are overlapped so better that will product higher inlier rate. Besides, compared to the total set which is with too complex scene, those underlying scenes out of the whole scene are more robust to be reconstructed and aligned in their own camera coordinate systems. To generate such subsets, we cannot resort to clustering algorithm like k-mean [5] to automatically cluster a number of photo sets because wrong matches based on visual similarity may occur due to the diversity. Meanwhile, although photos may be associated with meaningful textual tags, there is still no warrantee of that photos with the same salient phrase are visually relevant due to the tag noise [6].

In this paper, We build a search infrastructure to generate the subsets of photos with respect to underlying scenes. We embed a scene alignment algorithm using flow into the infrastructure, in particular using the SIFT flow [7] algorithm as we choose SIFT features to perform feature matching. we consider the photos nearby a place as an image database, given a photo that needs to be estimated its view direction, we find the best matched photos with respect to the given photo by searching in the database, thus generate a set of photos with significant overlaps. Note that a photo may occurs in multiple scenes, we choose the scene where that photo obtains the highest inlier rate when reconstructing the scene.

For each scene's reconstruction, we use RANSAC based normalized 8-point algorithm [9] to estimate the camera rotation, translation for simplicity. Although there are more robust algorithms can product higher inlier rate, such as [8], cost too expensive computation for large set of photos. Unfortunately, there is always placing a large amount of photos nearby a place on the world map. We do not perform ground evaluation because it is so difficult to get the ground truth for that whether a photo is exactly shot to the target place. Instead of we developed an online system for world exploration to perform the user study on our method.

2 The Methodology

Our goal is to find a set of matched photos for each given query example, facilitating the estimation of the camera rotation, translation. We describe our method in three steps as follows.

2.1 Search Infrastructure

Because objects present in photos are in different spatial location and captured at different scale in the case of Google Maps, we believe that the conventional methods for building a CBIR (query by example) system are not sufficient in effectiveness to return the most relevant results at most. To design a more robust CBIR, we search the photos by computing alignment energy using the flow estimation method to align photos in the same scene, where a photo example is aligned to its k-nearest neighbors in the photo collection. Since the use of SIFT features gives birth to robust matching across diversity, we employ the SIFT flow [7] algorithm to align them.

We introduce the search infrastructure as follows. We first extract SIFT features of all photos in the database, and index them using k-d tree algorithm. Given a query example of an underlying scene, we aim to find a set of matched photos from the database of a place to reconstruct the scene. The search results is returned by SIFT-based matching, and enhanced by scene alignment. In our method, we adopt k-nearest neighbors to search, and k=50 is used in our case, although other values also can be used. Figure 1 shows the top 10 nearest

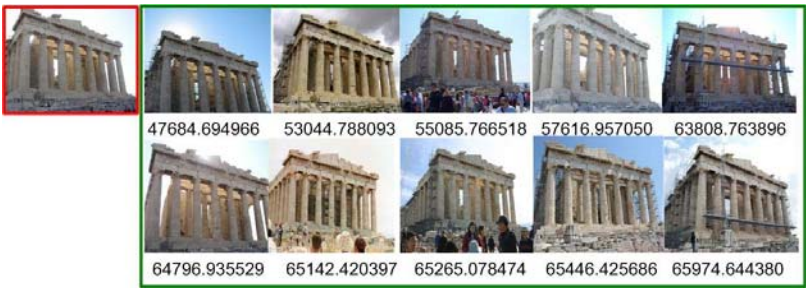

Fig. 1. An example of the search results returned by our search paradigm

neighbors of a query (marked with most left box). The number below each photo is the minimum energy of alignment obtained by the SIFT flow algorithm, which is used to rank the search results. As can be seen, this set of photos is promising to robustly estimate the extrinsic parameters of the corresponding cameras.

2.2 Estimation of Rotation, Translation

For each set of photos, we estimate the camera rotation, translation by using normalized 8-point algorithm to estimat the fundamental matrices, and apply RANSAC to detect the inliers for robust estimation of each fundamental matrix. The steps are detailed below.

1. For each pair of photos consists of p and $p^{'}$.
2. Determine the correspondences set C by SIFT features matching between p and $p^{'}$.
3. Randomly choose 8 correspondences and compute an initial fundamental matrix $\mathbf{F}$ using normalized 8-point algorithm, apply RANSAC to detect outliers, and determine inlier rate induced from current $\mathbf{F}$, if the appropriate inlier rate is achieved, the current $\mathbf{F}$ is the robust one, else repeat step 3. The $\mathbf{F}$ can be written as:

$$\mathbf{F} := \mathbf{K}^{'-1}\mathbf{TRK}^{-1} \tag{1}$$

   where $\mathbf{T}$ is the translation vector and $\mathbf{R}$ is a 3×3 rotation matrix.

Using a pinhole camera model, the 3D view direction V_p of the photo p can be obtained as $V_p = \mathbf{R}^{'} * [0\ 0\ -1]^{'}$, where $'$ indicates the transpose of a matrix or vector. We retain the x and y components of V_p as the 2D view direction of p. The 3D position is $-R' * T$, so that the 2D position is the remains of x and y components too.

2.3 Geo-registration of View Direction

Assume the geographic metadata is correct, we will use it expressed as latitude and longitude coordinate as the photo's position on the map. In this section, we begin geo-register the views of photos on the map by using the metadata and the whole procedure is illustrated in Figure. 3. In a reconstructed scene, for each pair of photos denoted as p and i, their coordinates and directions are denoted as L_{wcs_p} and L_{wcs_i}, V_{wcs_p} and V_{wcs_i} respectively. We clean the geographic metadata and get the GPS coordinates in terms of latitude and longitude as the photo's position on the map, which are marked as L_{geo_p} and L_{geo_i} in the figure. Since the angel demonstrated as ∂_p between the vector representing view direction and the vector linking two positions in the specific coordinate system is fixed. So we can register the directions in the camera coordinate system *wcs* onto the geographic coordinate system *geo* referenced by the fixed angels.

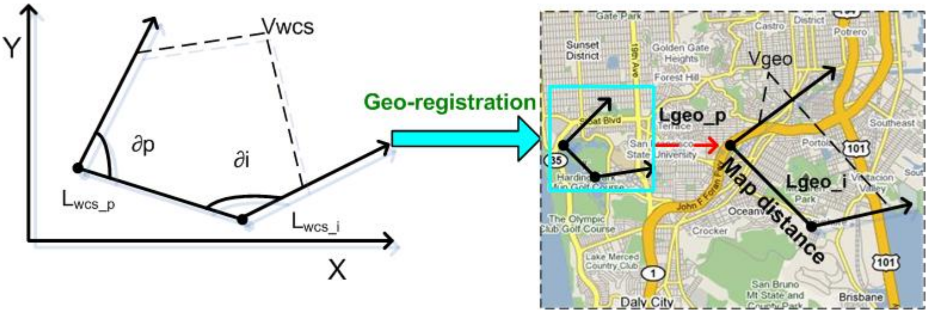

Fig. 2. Procedure of view direction's geo-registration

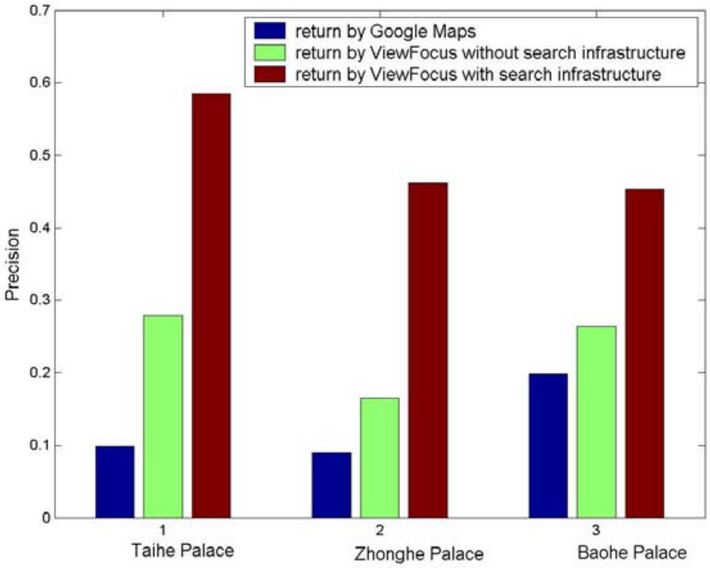

Fig. 3. Comparisons of precision evaluated on *ViewFocus*

3 User Study

Since it is difficult to obtain the ground truth about which photo nearby a place is exactly shot to the target region. We use a developed online system [10] named *ViewFocus* to evaluate our method by user study. This system allows users draw a rectangle on the map interface, then it returns the photos with view direction pointing to the enclosed region. Therefore, we can manually examine how many precise photos are returned so that to perform the user study. We use *precision* as the evaluation metrics. The settings of the evaluation are designed as follows: (a) examine how many related photos nearby the target regions on Google Maps; (b) examine how many related photos returned by the system using view direction filtering without the search infrastructure; and (c) examine how many related photos returned by the system using view direction filtering with the search infrastructure. In order to be more be judgeable about the photos' view directions, we select three famous buildings including "Taihe Palace", "Baohe Palace" and "Zhonghe Palace" in "Forbidden City" in Beijing, China. Therefore, assessors are more easy to justify whether the photos are shot to the target regions. We set the radius of the range centered at the target region to 100-meter in our evaluation. Therefore, there are 312, 383, 365 photos nearby these three buildings respectively. The comparisons of precision are presented in Figure 3. As can be seen, our method outperforms the others.

4 Conclusion

We presented a method to estimate the pose of the world's photos. We proposed a novel search infrastructure based on SIFT flow to generate number of sets of photos for scene reconstruction. Finally, we demonstrated the effectiveness by user study on the method proposed about the view direction and position of a photo.

References

1. Schaffalitzky, F., Zisserman, A.: Multi-view Matching for Unordered Image Sets, or How Do I Organize My Holiday Snaps? In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 414–431. Springer, Heidelberg (2002)
2. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Morgan Kaufmann, San Francisco (1987)
3. Luong, Q., Faugeras, O.: The fundamental matrix: Theory, algorithms and stability analysis. Int. J. Comput. Vision 17(1), 43–75 (1996)
4. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vision 60(6), 91–100 (2004)
5. Kanunqo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An Efficient k-Means Clustering Algorithm: Analysis and Implementation. IEEE Trans. Pattern Anal. Math. Intell. 24(7), 881–892 (2002)
6. Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.-T.: NUS-WIDE: A Real-World Web Image Database from National University of Singapore. In: Proc. of ACM Conf. on Image and Video Retrieval, CIVR 2009 (2009)
7. Liu, C., Yuen, J., Torralba, A., Sivic, J.: SIFT Flow: Dense Correspondence across Different Scenes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 28–42. Springer, Heidelberg (2008)
8. Urfalioḡlu, O.: Robust Estimation of Camera Rotation, Translation and Focal Length at High Outlier Rates. In: 1st Canadian Conference on Computer and Robot Vision (CRV 2004), pp. 464–471. IEEE Computer Society, Los Alamitos (2004)
9. Hartley, R.: In defence of the eight-point algorithm. IEEE Trans. Pattern Anal. Math. Intell. 19(6), 580–593 (1997)
10. Luo, Z., Li, H., Tang, J., Hong, R., Chua, T.-S.: ViewFocus: Explore Places of Interests on Google Maps Using Photos with View Direction Filtering. To appear in ACM Multimedia 2009 (2009)