OPTIMAL SELECTION OF SUBSET OF IMAGES WITH HIGHEST INTRA-CLASS SIMILARITY FOR 3D SCENE RECONSTRUCTION

Mahdi Salarian Rashid Ansari

University of Illinois at Chicago
Electrical and Computer Engineering Department
Chicago, Illinois 60607, United States

ABSTRACT

Finding the accurate location of a mobile device based on images it acquires usually requires applying structure from motion (SFM) for 3D camera position reconstruction. Since the convergence of SFM depends on effectively selecting among the multiple retrieved images, we propose an optimization framework to do make the selection using the criterion of the highest intra-class similarity among images returned from retrieval pipeline. The selection process should consider only images with distinct GPS-tags. The selected images along with the query can be used to reconstruct a 3D scene and obtain relative camera positions. Experimental results demonstrate our method achieves a higher convergence rate in the SFM processing.

Index Terms— Image-based localization, BOF, Retrieval.

1. INTRODUCTION

Finding the accurate location of a camera using an image it captures requires a search over a very large GPS-referenced image dataset collected from social sharing websites like Flickr [1] or services such as Google Street View (GSV) using image retrieval methods. The key tools employed in these methods are features such as SIFT [2] and SURF [3]. Although those features are powerful, the performance degrades with increasing size of the database, reducing the chances of finding the correct match. This can be overcome by having prior information about the approximate coordinates which can be used to narrow the search space down [4], [5], [6].

The process of finding the location does not end with finding the best match since retrieval engine often returns multiple images which are likely re-ranked by geometry verification. To acquire higher accuracy in estimating the location, methods based on SFM such as [7], [8], [9], and [10] can be used. Those methods utilize multiple Structure From Motion (SFM) to estimate camera locations. In order to avoid the time complexity of multiple SFM, we propose to use four relevant candidate images with distinct GPS-tags to be fed to SFM processing. This allow us to compute the query location by a closed-form solution and recover the query's real-world location. **Our contribution in this work is as follows:**

We propose a method to optimally select a subset of images from retrieved candidates with the highest intra-class similarity and distinct GPS-tags to increase the convergence rate of SFM. We note that it is critical to use distinct GPS-tags to perform proper post-processing after SFM for computing the transformation between camera-referenced coordinate and the real-world coordinate which is beyond our vision in this article.

In order to consider query features, we introduce a special similarity measure that takes into account those features that are common to all pairs of selected images and that are shared with the query as well. We show that our proposed method leads to higher convergence rate.

The rest of the paper is organized as follows. In the next section the problem for the optimal selection of images is formulated. Then, in Section III the method for solving the optimization problem is described and its implementation is discussed. Section IV demonstrate how our proposed method improves the performance in terms of convergence rate which is crucial for computing query camera location.

2. PROBLEM FORMULATION FOR OPTIMAL SELECTION OF IMAGES FOR SFM

In this section we demonstrate how the framework for the problem of optimal subset selection of images is formulated. After briefly describing the method used for image retrieval to obtain N matching images, we discuss about our proposed algorithm to optimally choose subset of k images to be used in SFM processing. Typically N may range between 10-50 whereas the choice of k is four in our work.

2.1. Retrieval of N Images

We first obtain N images that best match a query image. For this purpose several image retrieval methods may be employed. The main component of most image retrieval methods is the Bag Of Features (BOF) technique. In this approach,

each image is represented with a vector containing the occurrence frequency of features (visual words).

Let η be the number of visual words and $F_q = [f_1^q f_2^q, ..., f_\eta^q]$ and $F_{db} = [f_1^{db} f_2^{db}, ..., f_\eta^{db}]$ the frequency of visual words $w_1, w_2, ..., w_\eta$ for query and dataset image respectively. Each element of the f_j^q or f_j^{db} is the number of times the feature descriptors of the query or a dataset image has been assigned to visual word w_j. In order to damp the effect of visual words with higher frequency, algorithms such as Adaptive Assignment [11] or burstiness [12] can be used to improve regular BOF. Now the similarity between the query and a given dataset image (vectors) can be computed by calculating the product of the corresponding normalized vectors.

The final common step in most of image retrieval algorithms is to apply geometry verification based on RANSAC to re-rank the limited number of candidates based on number of inlier features. To go forward and estimate the actual position of the query, we need four candidates with distinct GPS tags. So another step should be applied to return images with distinct GPS tag and highest intra-class similarity. The process for selecting best candidates is discussed in next sub-section.

2.2. Optimum Selection of the Best k Retrieved Images

Suppose N images with location coordinates $g_i, i = 1, ..., N$ are selected after re-ranking. The simplest way to select k images is to find images with distinct GPS tags and select k images with the highest number of inliers. Those images are not necessarily the best choices for the the SFM processing since only the number of pairwise matches (inliers) between the query and each candidate is taken into account but not the number of matched features between each pair of candidates. It is important to note that a set of candidates is the best choice when each member of this set shares the highest number of common features with the other members. In our case, while multiple images per location exist, there is a need for a method to select the best set in a way that each member of the set has highest consensus on common features with other members as well as with the query image. The solution is facilitated by defining pairwise dissimilarity measure, w_{ij}, between distinct image i and j. An undirected graph $G = (V, E, w)$ with vertices $V = 1, 2, ..., N$ corresponding to image $I_1, I_2, ..., I_N$ with location $g_1, g_2, ..., g_N$, edges E, and weights w can then be created. By this definition, a more similar pair of images are going to have a lower w_{ij}. Now the problem is to find subset $G^\star = (V^\star, E^\star, w^\star)$, $V^\star \subset V$, $E^\star \subset E$, with k, $k < N$, vertices that minimize the total weights:

$$V^{k\star} = \underset{V^k \subset V}{\operatorname{argmin}} \sum_{\substack{i,j \in V^k \\ g_i \neq g_j \\ i \neq j}} w_{ij} \tag{1}$$

Here V can be partitioned into clusters with distinct GPS-tags. We now devise a solution to the problem of optimal selection of k images using the framework just described.

3. IMPLEMENTING SOLUTION TO OPTIMAL IMAGE SELECTION FOR SFM

The optimization problem of finding a subset from a set has been studied extensively during recent years [13], [14]. Since there are likely to be multiple images per location, the algorithm should only select one image per location. We therefore employ the General Minimum Clique Problem (GMCP) to select one among nodes with identical GPS-tag to acquire candidates with distinct GPS-tags. In the following subsection we show how our problem is solved by GMCP.

3.1. Candidates Selection By GMCP

In order to solve our problem by GMCP, we start with N best images from the retrieval result with world coordinates (GPS-tag) g_j, $j \in \{1, ..., N\}$, not necessarily distinct. Let h be the total number of distinct location coordinates. Then candidates are grouped into clusters $\{V_1, ...V_h\}, h \leq N$ so that images in each cluster have an identical GPS tag. So an arbitrary cluster V_r contains different number of images associated with coordinate g_r. With this setting some clusters only contain a single image meaning that the retrieval returns only one image for that location. Also h may be more than k (k is preferably 4) which is beyond our needs for the next step. One possible solution is to keep the first $k = 4$ clusters and find images with the highest similarity. We choose to keep $h(> k)$ clusters and finally select k images with the highest score from the result of GMCP. In order to solve our problem, for each member of all clusters, a similarity measure between image $i \in V_x$ and $j \in V_y$ where $x \neq y$ should be calculated. Number of inliers between pair of images derived from geometry verification is a great similarity indicator. Until now, we only found number of inliers between the query and a limited number of candidates which was computationally expensive. Applying geometry verification between each pair of candidates does not meet our needs since it would require an unacceptable amount of time. In order to avoid this time complexity we can use vectors containing frequency of visual words of images as defined in section 2. It is also important to incorporate the query visual words in computing the similarity between two images. This is because images selected in this stage along with the query should be fed to SFM pipeline. So the similarity measure should taking into account those visual words that are common to two images as well as to the query image. We therefore introduce a query-contextualized image similarity measure. Suppose the vector of visual words for image I is represented by $F_I = \{f_1^I, f_2^I, ..., f_\eta^I\}$. In order to incorporate query visual words in computing similarity, the indexes of non-zero visual words of the query are extracted

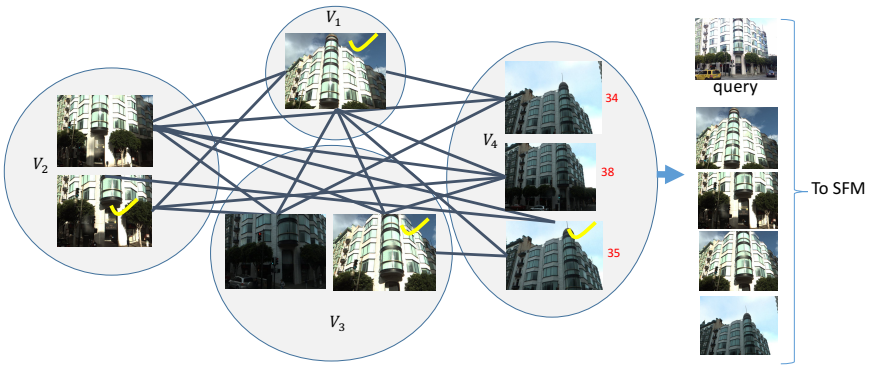

Fig. 1: Candidate selection by GMCP. Images with the same GPS-tag are placed in the same cluster. For cluster V_4 number of inliers between each member and the query is shown in red. For each cluster only one image marked by yellow check point returned by GMCP that maximizes the total weights (weights of edges are not shown in this picture). As shown for the cluster V_4, an images with a higher number of inliers is not selected.

and represented by $I^q_{nz} = \{u_1, ..., u_d\}$ which d is the number of non-zero visual words. Now the similarity between any pair of images i and j is defined by eq. 2.

$$\psi_{ij} = \sum_{k=1}^{d} \Delta(f^i_{u_k})\Delta(f^j_{u_k}) / (\sum_{k=1}^{d} \Delta^2(f^i_{u_k}))^{1/2} (\sum_{k=1}^{d} \Delta^2(f^j_{u_k}))^{1/2} \tag{2}$$

where, for $x \in \mathbb{R}$,

$$\Delta(x) = \begin{cases} 1 & \text{if } x \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Since $\Delta^2(x) = \Delta(x)$ the denominator in eq. 2 can be reduced to

$$(\sum_{k=1}^{d} \Delta(f^i_{u_k}) \sum_{k=1}^{d} \Delta(f^j_{u_k}))^{1/2} \tag{4}$$

This measure calculates the similarity between two images while taking into account the non-zero features of the query. The complexity of computing equation 2 is low since vectors are already available and summation is applied for the non-zero features of the query. A convenient measure of dissimilarity between image i and j can be defined by eq. 5.

$$w_{ij} = 1 - \psi_{ij} \tag{5}$$

The next step is to find a subgraph $G^\star = (V^\star, E^\star, w^\star)$ with nodes $V^\star = \{v^\star_1, ..., v^\star_h\} \subset V$ that only selects one node from each cluster, for instance $v^\star_1$ from V_1 and $v^\star_h$ from V_h, and subset of edges $E^\star \subset E$ that minimizes the total dissimilarity that for a feasible solution is:

$$T_{Dissimilarity}(V^\star) = \sum_{m=1}^{h} \sum_{l=m+1}^{h} w_{V^\star(m)V^\star(l)} \tag{6}$$

Fig. 1 shows the process of clustering images with only four clusters where the costs of edges are not shown. Only for the members of cluster one, V_4, the number of inliers between query and each member is shown in red. In this case, clusters contain different numbers of images. The result of GMCP is shown with yellow check marks. As is shown in cluster V_4, an image that has 35 inliers with query is selected. This result is different from that obtained with a method that only considers the number of inliers with the query. Without GMCP the best candidate for the cluster V_4 is the image with 38 inliers.

3.2. Generalized Minimum Clique Problem (GMCP)

Generalized Minimum Clique Problem (GMCP) can be used when the costs of edges are non-negative and graph is $|K|$-partite complete. While Minimum clique problem is based on nodes, GMCP substitutes nodes with cluster of nodes. In this problem nodes of a given graph are categorized in disconnected clusters. The goal is to find a subgraph with minimum cost or maximizing the score while only one node is selected from each cluster. On the other hand, each cluster introduces only one representive to the subset. This algorithm has been used recently in Computer Vision for multi-object tracking [15]. Suppose we are given a graph $G = (V, E, w)$ with nodes $V = \{v_1, ..., v_N\}$ and nodes are grouped into h sets of nodes called clusters (i.e. $V = V_1 \cup V_2 \cup ... \cup V_h$ and $V_x \cap V_y = \varnothing$ for all $x, y \in \{1, ..., h\}$ where $h \in \mathbb{Z}$: $1 \leq h \leq N$ with $x \neq y$ and a cost w_{ij} is considered for edge between nodes $i \in V_x$ and $j \in V_y$, for $x \neq y$. Now the objective is to find subgraph $G^\star = (V^\star, E^\star, w^\star)$ with nodes $V^\star = \{v^\star_1, ..., v^\star_h\} \subset V$ which is composed of only one node from each cluster and subset of edges $E^\star \subset E$ that minimizes the total edge costs. For such a problem GMCP can find a feasible solution with minimum cost which is in fact the total weights of all edges in $E^\star$. So based on formulation of our problem in section 3.1, GMCP can return the subset

Fig. 2: a) Query image. b) Images returned by retrieval pipeline for the query image. c) Images selected by proposed method based on GMCP. d) Images selected by finding images with highest number of inliers with query and distinct GPS-tag. Although images from two sets seem to be similar, only set returned by GMCP converged in the SFM pipeline

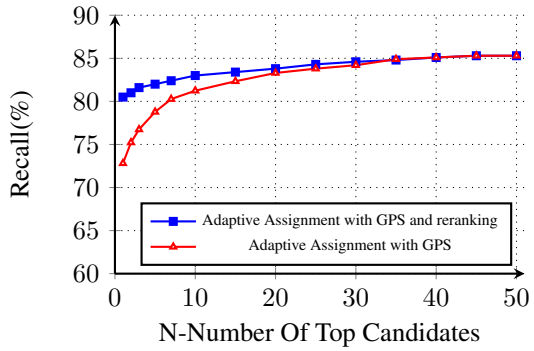

Fig. 3: Recall VS number of the candidates for San Francisco dataset. Limiting the search area by considering the query rough location from noisy GPS is used to improves the recall

with highest intra-cluster similarity which leads to improved convergence rate in the SFM step.

4. PERFORMANCE EVALUATION

In this research, we evaluated the performance of our method using online San Fransisco dataset from [16] containing more than one million images. The reason for using this dataset is that it contains more images per area which is necessary in our research which is based on four images with distinct GPS-tag in SFM process. The San Francisco dataset provides a set of 803 query images, usually taken from a pedestrian's perspective at street level. We also used Adaptive Assignment [11] while $\eta = 200k$ for the image retrieval engine. To assess the performance, *recall* as used in [16], [17], has been employed. To further improve recall rate, rough position and maximum GPS error are used for narrowing down the search space. Result is presented in Fig. 3.

We found that relevant images typically have more than 20 inliers. So candidate images with fewer than 20 inliers have been filtered out directly. From 803 original queries, our retrieval pipeline finds candidates which have at least 20 inliers for 453 queries . For 398 queries, more than four images are found. Although retrieval curves for $N = 50$ are shown, we have selected 15 images for the GMCP ($N = 15$). The reason is that the recall is almost flat for the $N > 15$. A subset of four images is then selected with two different approaches discussed in section 2.2. For queries for which the number of retrieved candidates is less than 15, all retrieved images proceed to the next step. Fig. 2 shows a query with multiple candidates returned from the retrieval pipeline and four images opted by the two approaches. Although images appear to be similar in both sets, the set returned by GMCP converged in SFM processing while the other did not.

For the 277 queries from 398, both approaches, returned identical subsets. Among those sets, 141 of them converges and produces 3D coordinates. For the reminding 121 queries we got different subsets with 42 convergences for the method based on finding distinct GPS tag and 61 convergences for the GMCP based approach. It is worth mentioning that GMCP based selection converged for all samples for which the distinct tag based method converged. Also it is important to note that we do not incur any significant increase in computational burden in our method with four images. This is because image selection based on GMCP with $N = 15$ nodes does not require a huge amount of computation and adds up less than 10% to the time required for retrieving and re-ranking images for an arbitrary query.

5. CONCLUSION

In this research, we propose a method to optimally select the best subset of images selected with the highest similarity to be used in reconstructing a 3D scene by using SFM. This method considers not only similarity between the query and a particular candidate, but all similarities between each pair of candidates. Experimental results show that our approach is able to achieve higher convergence rate in SFM process. Also we noticed that our proposed method will have higher convergence rate for a larger set of query images if the original database has more images per location and a higher degree of overlap between images from similar locations.

6. REFERENCES

[1] J. Sang, T. Mei, Y.-Q. Xu, C. Zhao, C. Xu, and S. Li, "Interaction design for mobile visual search," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1665–1676, 2013.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[4] J. Zhang, A. Hallquist, E. Liang, and A. Zakhor, "Location-based image retrieval for urban environments," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 3677–3680.

[5] M. Salarian, A. Manavella, and R. Ansari, "Accurate localization in dense urban area using google street view images," in *SAI Intelligent Systems Conference (IntelliSys), 2015*. IEEE, 2015, pp. 485–490.

[6] M. Salarian and R. Ansari, "Improved image retrieval for efficient localization in urban areas using location uncertainty data," in *IEEE International Symposium on Multimedia (ISM),*. IEEE, 2016.

[7] X. Xu, T. Mei, W. Zeng, N. Yu, and J. Luo, "Amigo: Accurate mobile image geotagging," in *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service*, ser. ICIMCS '12. New York, NY, USA: ACM, 2012, pp. 11–14. [Online]. Available: http://doi.acm.org/10.1145/2382336.2382340

[8] K. Vishal, J. C.V., and C. Visesh, "Accurate localization by fusing images and gps signals," *CVPR IEEE*, 2015.

[9] M. Salarian, N. Ileiv, and R. Ansari, "Accurate image based localization by applying sfm and coordinate system registration," in *Multimedia (ISM), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 189–192.

[10] A. Roshan Zamir, S. Ardeshir, and M. Shah, "Gps-tag refinement using random walks with an adaptive damping factor," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4280–4287.

[11] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 883–890.

[12] H. Jégou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1169–1176.

[13] N. Katoh, T. Ibaraki, and H. Mine, "An algorithm for finding k minimum spanning trees," *SIAM Journal on Computing*, vol. 10, no. 2, pp. 247–255, 1981.

[14] M. Fischetti, H. W. Hamacher, K. Jørnsten, and F. Maffioli, "Weighted k-cardinality trees: Complexity and polyhedral structure," *Networks*, vol. 24, no. 1, pp. 11–21, 1994.

[15] A. R. Zamir, A. Dehghan, and M. Shah, "Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 343–356.

[16] D. Chen, G. Baatz, K. Koser, S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, and M. Pollefeys, "City-scale landmark identification on mobile devices," *CVPR IEEE*, 2011.

[17] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt, "Image retrieval for image-based localization revisited." in *BMVC*, vol. 1, no. 2, 2012, p. 4.