An efficient method is proposed for refining GPS-acquired location coordinates in urban areas using camera images, Google Street View (GSV) and sensor parameters. The main goal is to compensate for GPS location imprecision in dense area of cities due to proximity to walls and buildings. Available methods for better localization often use visual information by using query images acquired with camera-equipped mobile devices and applying image retrieval techniques to find the closest match in a GPS-referenced image dataset. The search areas required for reliable search are about 1-2 sq. Km and the accuracy is typically 25-100 meters. Here we describe a method based on image retrieval where a reliable search can be confined to areas of 0.01 sq. Km and the accuracy in our experiments is less than 10 meters. To test our procedure we created a database by acquiring all Google Street View images close to what is seen by a pedestrian in a large region of downtown Chicago and saved all coordinates and orientation data to be used for confining our search region. Prior knowledge from approximate position of query image is leveraged to address complexity and accuracy issues of our search in a large scale geo-tagged dataset. One key aspect that differentiates our work is that it utilizes the sensor information of GPS SOS and the camera orientation in improving localization. Finally we demonstrate retrieval-based technique are less accurate in sparse open areas compared with purely GPS measurement. The effectiveness of our approach is discussed in detail and experimental results show improved performance when compared with regular approaches.

Index Terms— mobile device, LBS, Geotagged-images, TF-IDF, GSW

1. INTRODUCTION

The estimate of the location of a mobile device is valuable in a variety of different applications such as navigation and augmented reality. Accurate location is especially critical for blind or visually impaired pedestrians to safely navigate outdoors in cities. As a result accurate localization has recent been a very active area of research [***]. Even though traditional approaches that utilize the GPS module data or distance from cellular towers are useful for this performance, how satisfactory their performance is depends on the accuracy provided by the satellite navigation system. GPS information is usually satisfactory in many applications when the device has a clear view of sky to get the signal from at least 4 satellites. Using a GPS-equipped device carried by a pedestrian who is moving slowly on a street sidewalk close to tall walls and buildings, it is difficult to obtain accurate localization. This difficulty is generally most pronounced in dense region of cities. However the GPS errors of mobile phone are rarely greater than 100 meter [1]. Taking this into account, some researchers investigated approaches that use inertial sensors of the device to address GPS unreliability. Since accuracy is vital for applications in which people are highly dependent to Location-Based Services (LBS) technologies, such as the blind and visual impaired people, several research efforts have been directed at finding solutions to compensate GPS shortcomings. In order to have a better performance, these efforts have sought to use sensors such as gyro and accelerometer that are available in most of the mobile devices. It clearly desirable to exploit as much information from the sensors available in mobile devices. A large part of the effort to improve localization focused on using the camera that is an essential part of any smartphones and mobile devices. This effort addressed the possibility of getting accurate positions of a query image generated by the camera using image retrieval methods. Researchers investigated techniques for seeking the best match for a query image in a database of Geo-tagged images with accurate GPS coordinates. Some of these investigations relied on the use of datasets provided by researchers, while others used databases of images which are publicly available such as Google Street view and Flicker. Following the database search, the best match for a query image captured by the smartphone or mobile device is found, leading to better estimates of the coordinates of query. Existing image retrieval approaches have been found to be successful in getting matches especially when images have adequate texture. The key tools employed in these methods are based on scale-invariant feature such as SIFT [2], SURF [3] or DSIFT. Although these features are powerful the performance degrades with increasing size of the dataset, reducing the chances of finding the correct match. This can be overcome by having a prior information about the approximate coordinates which can be used to narrow down the search space. For example,
in [4] the database is split into overlapping regions for the search, or in [5] only a limited number of images around the approximate position provided by GPS is searched to find the best match. These methods are attractive as they are more suited for practical application since even clustering features extracted from city-scale dataset is very time consuming and needs a large memory. Relying purely on image retrieval techniques for outdoor navigation is not justifiable since we may use any available sensor data such as GPS available in new generation of phones. On the other hand considering the fact that the GPS error can be as large as 100 meters in some locations, query images generated by smartphone cameras can aid in refining localization where the GPS is not accurate enough. One piece of information that is very valuable in location estimate refinement is the uncertainty in GPS-based location that is accessible using the method described in [14].

**In our proposed method we exploit the accessed information about variable GPS error parameter** to adaptively limit the search area for image retrieval. Although a large body of research addresses the problem of localization using purely retrieval techniques, but our results show that those methods are often inferior to pure GPS location especially when we use public images from services like Google street view. Our method is also adaptive in the sense that it decides where is better to use GPS data by itself and to use retrieval-assisted refinement to achieve a better result. **One other key piece of accessible sensor information we use is the camera orientation** to limit the images to lie within a selected angle of view. The organization of this paper is as follow. In next section related work about determining device location using retrieval techniques is described. Then in section 3 our method for creating the dataset of images in a dense area of the city of Chicago is discussed. In section 4 our image retrieval algorithm and the proposed method for narrowing search space are discussed. Results on the performance of method for different data samples along with conclusions are presented in section 5.

**2. RELATED WORK**

Recent computer vision advances have made it possible to recognize natural and man-made objects in outdoor environment with sufficient reliability for many applications. A major application for this capability is using a massive number of Geo-tagged images on the internet to find the location of a query image [6]. A variety of methods have been proposed to do this. For instance, Retiary and Drummond [7] utilized an edge-based method to get street facades based on tree dimensional method. Other researches proposed modifications of the method by using an initializing step that is based on an accurate GPS antenna [8]. The most efficient and accurate approach is based on information such as scale-invariant features (SIFT) and its variations. After extracting and storing features for images in a database, features of a query image are determined and the best match is sought by searching stored features of images in the database using content-based image retrieval (CBIR) approaches. An effective approach that is widely used is Bag of Features (BOF) approach that is based on adapting text retrieval techniques to image retrieval, proposed by Zisserman and Sivic [9]. In this method all features descriptions are quantized to visual words with a clustering algorithm like k-means. An image is represented by a histogram of a number of visual words and each image in a database has its own histogram. Finding the best match, the histogram of a query image is compared with all histograms in database to infer which member of database closely matches the query. There are different measures for finding similarity such as the inner product of two BOF vectors. Although similarity measures can be calculated easily by using, for example, L1 distance, a widely used procedure in the inverted file [10] method which shows excellent performance. Some other researchers focused on clustering step to find an efficient quantization technique to assign each feature descriptor to a visual word. For example, Hierarchical k-means and Approximate k-means are preferred when we deal with a huge amount of data. Regular k-means is not suitable for huge number of images. Modifications such as soft assignment instead of hard assignment have been proposed to compensate incorrect assignment of a sample feature vector [11]. This is not necessarily the last step in the retrieval. Several methods that are based on image retrieval techniques select more just a single best matched candidate. The number of plausible candidates for each application is different and may vary from 10 to 100. An additional step called Homology verification is used to choose the best match among the plausible candidates. This step generally utilizes the popular algorithm of RANSAC [12] to find the best geometric match between query and plausible candidates with strong corresponding features. In fact this step compensates for the weakness of image retrieval schemes based on BOF where the geometric information of images is ignored. One distinctive work [13] proposes a method to use inertial sensor information and BOF to get more accurate result. Specifically, the wireless cellular information is inferred from the GPS-based location coordinates and this is used to limit the search to the cellular area. Such methods have shown reasonable performance. It should however be noted that the cellular area is much larger than the area corresponding to uncertainty in the GPS.

We emphasize that all the above-mentioned approaches do not exploit the accessible information about the **uncertainty in GPS coordinates** and the **camera orientation**.

![Figure 1: Place that our dataset](image)
3. CREATING THE GSV DATABASE
Available databases are not suited for our purpose due to the nature of the sensor data we use in our method. To create the database for investigating our method, Google Street View (GSV) service is used. This service covers most of US cities and more than 7 countries in 4 continents. Although it is free, but in case of making a dataset including thousands of images we had to seek permission from Google for downloading images. With this permit granted, utilizing Google street view API makes it easy to download image directly from a URL. Here first a Java script program is used to find coordinates that contains Google street images. This program extracts the coordinates and a radius of uncertainty and randomly selects different points and calls a function from Google map API to get a point that is closest to a GSV image. If this point has not been saved before, then it is added in the database. This procedure continues until the program does not find new new coordinates. After storing all possible coordinates, we generate a URL request using a Matlab-based program and send it to download images for each unique set of coordinates [5]. The URL request includes latitude, longitude, pitch, yaw, heading and field-of-view (FOV) is sent to get each image. Since our target is accurate localization, pitch and FOV are assigned values of 10 and zero to get images that are similar to the viewpoint of a pedestrian. This may be customized to a pedestrian if needed. Also for each set of coordinates we considered 12 images separated by 30 degree by changing heading. The dataset contains coordinates and orientation of all images and these coordinates have high accuracy because Google vehicles are equipped with high-precision instrumentation. Using the accurate coordinates and orientation parameters such as pitch and yaw, we sought to determine whether the image from camera and Google street view are similar. Figure 1 shows a region that is a dense part of downtown Chicago. In this region, shown with orange, are almost 20k panoramic images. With each panoramic image consisting to 12 30-degree views, the total number of images in that region is roughly 20k×12=240k images. Because of limitation in capacity of the server we used, a portion of that region containing 24k images is used as our dataset. It is clear that a city scale dataset can be split into some overlapping regions as in [4]. In our study, the size of data set is not an issue when prior knowledge from GPS and sensor are taken into consideration. Experimental results show that the computational cost for filtering images by approximate GPS coordinates and heading is lower than the total cost for retrieval.

4. PROPOSED METHOD
There are variety of researches for finding similarity between images. Central to almost all approaches is feature extraction techniques. In recent years, most of the research has largely focused on SIFT and its variants, which have proved to be effective in image retrieval. We adopted the dense SIFT method in our work. We also considered the popular BOW and inverted file method as visual recognition engine for search. In BOW each image is described by a vector of visual words. Steps for implementation of BOW is as follow.

a) Find dense SIFT feature for all images in a dataset
b) Cluster features
Hierarchical k-means (HKM) is used with branching factor of 10 and depth of 5. So we have 10^5 clusters (visual words).
c) Find closest visual word (cluster center) for each feature in images. Represent each image by a vector showing the number of each visual word in it.
d) Find best match between 2 images by comparing their vectors. This can be done by calculating distance between 2 vectors.

Although images can be represented by visual words, the importance of the words varies. Different schemes have been proposed to consider the weight for each visual word. The most popular one is TF-IDF whitening scheme and Inverted file. TF (term frequency) considers the importance of each visual word (term) in each image while IDF (inverse document frequency) reflects importance of each term in whole dataset. We considered TF-IDF scoring scheme in (1): where

\[ w_{ij} = tf_{ij} \times \log\left(\frac{N}{df_i}\right) \]

1. Formula for calculating TF-IDF weighting
N is the number of images in dataset, \( df_i \) is number of images containing \( t_i \) and \( tf_{ij} \) is the frequency of term \( i \) in image \( j \).
RESULTS AND CONCLUSION

Currently GPS that is widely available for use by civilians in the USA is not accurate enough for navigation especially in dense area of cities. People who need LBS are adversely affected by this limitation. In this research a method based on image retrieval techniques is proposed to improve localization. Noting that useful sensors are embedded in most of Mobile devices, a new solution to localization based on restriction on search area by using DOP extracted directly from GPS is used to not only reduce the running time, but improve the performance of visual recognition engine. For the purpose of limiting the search area we utilized DOF parameter calculated by GPS data together with the camera angle. Results of performance are shown in Table 1 for 100 query images and for the three scenarios of search in whole data set, local search using GPS uncertainty, and local search combined with camera angle (FOV). Although image search is successful in most of cases but the performance is not perfect. To improve our results we will exploit the GSV image data and camera location relative to it to provide geometric corrections to improve localization. We will also explore alternatives such as modified schemes for TF-IDF because its weighting directly affects the result.

| Search in whole data set | Local search | Local search and narrowing angle (K = 30°) |
|--------------------------|--------------|------------------------------------------|
| Voting                   | %80          | %96                                      |
| TFIDF                    | %56          | %92                                      |

Table 1: Performance of system with different scenario
11. REFERENCES

[1] G. Schroth, R. Huitl, D. Chen, M. Abu-Alqumsan, A. Al-Nuaimi, and E. Steinbach. Mobile visual location recognition. IEEE Signal Processing Magazine, 28(4):77–89, 2011.

[2] D. Lowe. Distinctive image features from scale-invariant key points, International Journal of Computer Vision, 60(2):91–110, 2004.

[3] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding (CVIU), 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, 18th IEEE conference in image processing, 2011.

[5] arXiv:1412.8496v1.

[6] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), volume 0, pages 1–7, 2007.

[7] G. Reitmayr and T. W. Drummond. Going out: Robust tracking for outdoor augmented reality. In Proc. Intl. Symp. on Mixed and Augmented Reality (ISMAR), pages 109–118, 2006.

[8] G. Reitmayr and T. Drummond. Initialisation for visual tracking in urban environments. In Proc. Intl. Symp. on Mixed and Augmented Reality (ISMAR), pages 161–160, 2007.

[9] J. Sivic and A. Zisserman, “Video Google: A text retrieval approach to object matching in videos,” in Proc. IEEE Int. Conf. Computer Vision, Nice, France, Oct. 2003, pp. 1470–1477.

[10] I. Witten, A. Moffat, and T. Bell, Managing Gigabytes: Compressing and Indexing Documents and Images. San Mateo, CA: Morgan Kaufmann, 1999.

[11] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman Improving Particular Object Retrieval in Large Scale Image Databases, IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008.

[12] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6):381–395,1981.

[13] Tao Guan, Yunfeng He, Juan Gao, Jianzhong Yang, and Junqing Yu, On-Device Mobile Visual Location Recognition by Integrating Vision and Inertial Sensors, IEEE transactions on multimedia, vol. 15, No. 7, November 2013.

[14] M. Sharif a, A. Stein a, Ernst M. Schetselaar, Integrated approach to predict confidence of GPS measurement.