EXPLOITING PRIOR KNOWLEDGE IN MOBILE VISUAL LOCATION RECOGNITION

G. Schroth, R. Huitl, M. Abu-Alqumsan, F. Schweiger, E. Steinbach

Institute for Media Technology, Technische Universität München, Munich
{schroth, huitl, moh.alqumsan, florian.schweiger, eckehard.steinbach}@tum.de

ABSTRACT

Mobile visual location recognition needs to be performed in real-time for location based services to be perceived as useful. We describe and validate an approach that eliminates the network delay by preloading partial visual vocabularies to the mobile device. Retrieval performance is significantly increased by composing partial vocabularies based on the uncertainty about the location of the client. This way, prior knowledge is efficiently integrated into the matching process. Based on compressed feature sets, infrequently uploaded from the mobile device, the server estimates the client location and its uncertainty by fusing consecutive query results using a particle filter.

Index Terms— Image Retrieval, Mobile Visual Location Recognition, Bag-of-Features,

1. INTRODUCTION

Mobile visual location recognition aims at continuously deriving the pose of a mobile device by matching the current camera view to a typically large set of reference views. As GPS is hardly available in urban canyons and indoors, visual location recognition enables location based services (LBS) in these densely populated areas without the need for complex infrastructure. The task of matching the current recordings to a large reference database like Google Street View [1] or Mircosoft StreetSide [2] is known as content based image retrieval (CBIR), which has been an active area of research for years. Most state-of-the-art retrieval systems base upon the bag-of-features (BoF) approach [3, 4]. Feature descriptors are quantized to so called visual words, e.g., using k-means, to represent an image by a sparse visual word frequency vector. The visually most similar images can be efficiently determined by computing the pairwise distance between query and database BoF-vectors using an inverted file scheme.

The application of CBIR to mobile location recognition raises new challenges when compared to mobile product recognition [5, 6]. A severe amount of clutter and dynamic objects like cars, pedestrians, and advertisements can be found in both query and reference images. Feature bundling techniques [7] may be required to include local geometric properties into the matching process to increase distinctiveness. Due to complex 3-dimensional scenes and occlusions, the appearance of a location may change significantly between different perspectives. Finally, the client-server architecture has to be designed in a way that minimizes both the computational load assigned to the mobile device as well as the delay induced by the communication with the server. Due to the rapidly changing field of view of the mobile device caused by camera panning, it is essential to achieve close to real-time query results for LBS to be perceived as useful. On the other hand, weak prior knowledge on the location derived from cell-ids or the last GPS localization can be always assumed. Most importantly, however, successive query frames are recorded in their respective vicinity.

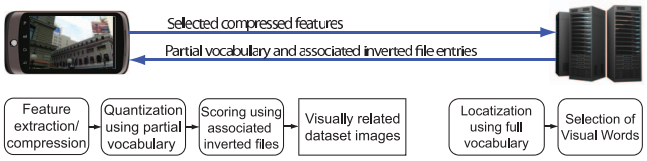

Fig. 1. System overview illustrating the use of partial visual vocabularies at the client (adapted from [6]).

Based on this assumption we aim not only at reducing the computational complexity at the mobile device but also eliminate the network delay by preloading selected reference data to the device. By sending the most relevant data to the client using the typically 5 times faster downlink, the client can localize itself within a limited area without any communication with the server. Please note that the general idea has already been proposed in our overview paper on mobile location recognition in [6]. In this paper we validate the approach and evaluate its performance. Additionally, we propose a particle filter based approach to estimate the client location at the server using infrequently uploaded query feature sets. This not only reduces the amount of data to be sent to the client but also significantly increases the retrieval precision as prior knowledge is efficiently integrated into the matching process.

The remainder of the paper is structured as follows. In the following section we discuss state-of-the-art approaches on compressing the data to be transmitted to the server as well as systems to integrate prior knowledge on the location. In Section 3 we introduce the concept of partial visual vocabularies. The selection of visual words to form partial vocabularies is discussed in Sections 3.1 and 3.2. To evaluate the properties of the approach, the results of real world experiments are described in Section 4. Finally, we conclude the paper with an outlook to future work.

2. RELATED WORK

The delay caused by the communication between the client and the server using a typical 3G network has been studied in [5]. The transmission of 10 kB using the uplink of an indoor 3G connection takes about 5 to 8 seconds on average depending on the signal quality. This does not include timeouts, which happen in about 6% of the connections.

Several approaches have been proposed to minimize the amount of data to be transmitted to the server. Chandrasekhar et al. [8] introduced a Compressed Histogram of Gradients (CHoG) descriptor, which exploits the non-uniform distribution of gradients to describe an image patch using only 60 bits. Hence, an image can be represented by 3-4 kB requiring approximately 3 seconds to be uploaded.

Chen et al. proposed to quantize the features on the mobile device into visual words and to compress the sparse visual word frequency vector, which allows for a 5-times rate reduction with respect to CHoG [9]. This, however, requires downloading the quantization

structure to the mobile device, which would result in a time consuming initialization or to use a generalized quantizer which would lead to inferior the retrieval performance. Further, the quantization into a vocabulary of about one million visual words on the mobile device adds additional complexity and hence delay to the overall retrieval.

Recently, Ji et al. [10] proposed to integrate weak prior knowledge on the location to adapt the allowed set of visual word indices to be sent to the server. Based on a generalized visual vocabulary the subset is selected based on ground truth data of previously selected distinctive landmarks. To distinguish among the approximately 60 landmarks per city, less than 32 bits are needed in [10] to represent an image. This approach, however, can be hardly adopted to continuous location recognition at arbitrary query locations as representative exemplary query images can not be generated. Further, even with very limited amount of data to be transmitted to the server, the roundtrip delay of about 0.3 to 0.9 s can hardly be avoided in this client-server architecture.

3. PARTIAL VISUAL VOCABULARIES

In contrast to prior work, we exploit the typically 5 times faster downlink to preload selected reference data to the mobile device. This allows us to perform the localization within a limited area without waiting for responses from the server.

As shown in Fig. 1, the features of every kth frame are transmitted to the server. To minimize the time required to upload the data, compressed CHoG [8] descriptors are used. Multiple vocabularies, each covering an area of about 5 km^2 comprising approximately 6000 panoramas, are available at the server. Based on the weak prior knowledge the most suitable vocabulary is selected and used to quantize the uploaded features into its visual words (typically one million) using approximate k-means (AKM) [4]. As 500 CHoG descriptors can be uploaded in about 3 seconds, the server can determine the visually most similar panorama and thus the location of the mobile device typically every 90th frame assuming 30 fps. With these periodic location estimates, we can limit the reference data required at the client to localize itself until the next server-based estimate to a fraction of the full vocabulary and inverted file system.

If we consider $F = \{f_1, f_2, ..., f_N\}$ to be the set of features of one query frame and $V = \{v_1, v_2, ..., v_L\}$ to be the set of visual words (i.e., the full vocabulary), the quantization function $q_V(f_i) = v_j$ assigns each feature f_i to the closest visual word v_j in the full vocabulary V. This is typically done by performing a nearest neighbor search among all visual words. The subset of visual words, which represents a particular video frame, is defined as $Q(F|V) = \{v = q_V(f) | f \in F\} = V_F \subseteq V$. The result of exact nearest neighbor quantization remains unchanged if only the subset of visual words representing the frame itself $V_F = Q(F|V)$ is used instead of the full vocabulary:

$$Q(F|V_F) = Q(F|V) \tag{1}$$

Hence, only this part of the full vocabulary needs to be available at the client to obtain the same results. However, this equation only holds for a specific set of features F. A partial vocabulary V_F would need to be sent to the client for each frame. Ideally, we would like to identify a partial vocabulary that includes the correct visual words to process multiple consecutive frames without the need to know their features a priori. Since V_F can be extended by other subsets of the full vocabulary ($S \subseteq V$) without changing the quantization result, as shown in Eq. 2, we can use partial vocabularies at the client that have a sufficiently high probability of including V_F.

$$Q(F|V_F \cup S) = Q(F|V_F) = Q(F|V) \tag{2}$$

To limit the amount of data to be transferred to the client, we seek for the smallest partial vocabulary that includes the unknown V_F for the next video frame(s) with high probability.

3.1. Composing the partial vocabularies

While the field of view of the mobile device may change rapidly due to the varying user attention, the location typically changes only at a walking pace of about 1.2 m/s. Thus, visual words representing successive frames are found with a high probability in the same reference image, i.e., a panorama, at the corresponding location, which is periodically retrieved at the server every kth frame. As the distance between neighboring panoramas ranges between 12 to 17 m, the visual words of two panoramas are sufficient for the client to localize itself for about 10 s. Thus, the set S in Eq. 2 is composed of the visual words from the top ranked K panoramas retrieved by the server and their neighboring panoramas within a radius R. While this approach does not ensure to include the complete sets V_F of the next video frames in the partial vocabulary, this interestingly improves performance as it effectively integrates prior knowledge on the location into the quantization process at the client. Features can only be quantized into visual words located in the area of the current location uncertainty. Further, due to the small size of partial vocabularies, the quantization of features can be performed at high rates on the mobile device, achieving 10 fps including feature extraction on a state-of-the-art phone.

The number of visual words to be downloaded depends on the currently available data rate and the time until the next localization update at the server (step size). As the visual words and their associated inverted files are progressively downloaded and added to the quantization structure, the visual words included in the closest panoramas are transmitted first. The radius R is set according to the expected distance travelled until the next server update. Less time has to be spent on downloading the visual words from neighboring panoramas as they typically share about 20% of their visual words.

Using a partial vocabulary composed of the top ranked K panoramas and their neighbors within a radius R yields very good results but exploits only the information obtained from a single set of features uploaded to the server. The area of location uncertainty can be better estimated using a particle filter which fuses the information obtained from consecutive query sets.

3.2. Particle Filter based partial vocabularies

By tracking multiple particles, i.e., hypotheses, on the pose and velocity of the mobile device at the server, we can fuse consecutive retrieval results by exploiting a constant velocity model. As the client can be assumed to have a zero centered Gaussian distributed acceleration, a given maximum velocity and rate of turn, the motion model allows us to predict possible locations until the next retrieval update becomes available. Based on the state vectors of the last set of particles, new hypotheses are generated, which are samples of the probability density function (pdf) of the estimated client location. In addition, single particles are added at the location of the last retrieved panoramas. This allows us to reduce the number of particles and to account for distant but consistent retrieval results.

Every time a new set of features arrives at the server, the particles are weighted according to the probability that the retrieval results are obtained at their location. Using a ground truth dataset we learned a pdf that specifies the probability of retrieving a given panorama at a given distance. This pdf is specific to a given retrieval pipeline. As subsequent updates occur typically every 5 to 25 seconds, the temporal correlation of the noise in the retrieval results is small. Thus,

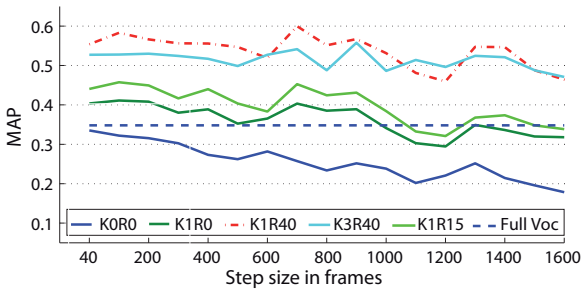

Fig. 2. MAP scores for multiple partial vocabulary configurations over the update rate (step size).

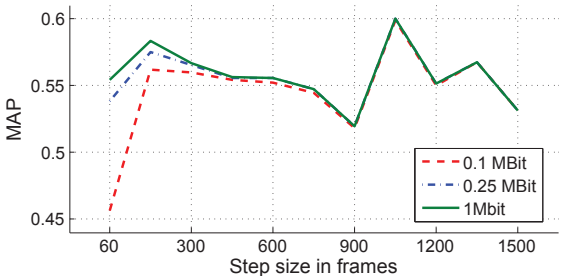

Fig. 3. MAP score of the K1R40 partial vocabulary configuration at three different data rates over the update rate (step size).

false retrieval results are ruled out by the filter as their associated locations do not comply with the motion model. Based on this low complexity particle filter, we can avoid time consuming geometric verification of the retrieval results at the server.

Based on the resulting pdf over the estimated client location we compose the partial vocabulary of visual words that best cover this area. As we also estimate the direction of motion as well as the velocity, significantly fewer visual words have to be transmitted to achieve the same or better retrieval results at the client.

Further, by utilizing database statistics we identify those visual words that provide the maximum amount of information to distinguish individual locations [11] and, at the same time, are most frequently found at a specific location. This allows us to reduce the number of visual words representing a panorama by at least a third at hardly any loss in retrieval performance. This way, less than 1000 visual words are sufficient to represent a full panorama.

Visual words are represented by compressed CHoG descriptors (60 bit) and inverted file entries are encoded differentially [9] to further reduce the amount of data to be downloaded. On average a total of 140 bit per new visual word and associated inverted file entries (typically 6 in a database of 6 thousand images (5 km^2) and 1 million visual words) has to be downloaded. However, as more and more visual words are already available at the client, only their id (20 bits) has to be transmitted to form the partial vocabulary at the client. The data transmission will be discussed in more detail in Section 4.

4. EXPERIMENTAL EVALUATION

To evaluate the performance of our approach, we applied it to a realistic scenario where the client is travelling along a 650 m long track in downtown San Francisco recording 30 fps with a resolution of 840x480 pixels. The velocity ranges between 0.5 to 1.9 m/s. Recordings are impaired by motion blur and a severe amount of clutter like cars and foliage. Prior knowledge derived from Cell-IDs allows us to select one of the overlapping 5 km^2 subareas at the server which includes 6 thousand Google Street View panoramas with an inter panorama distance ranging from 12 to 17 m.

Every k frames (k is called step-size) features are sent from the mobile device to the server. We extract MSER features [12] from the least blurry frames within the last 10% of the step size. Based on the retrieval results at the server, which employs an AKM [4] with 1 million visual words, the partial vocabularies are formed as described in Section 3.1. The client performs feature extraction and retrieval in the partial vocabulary at 10 fps to continuously determine its pose without interference by the server. Query features are matched to the visual words using a forest of four extremely randomized kd-trees [13] that have to be only rarely updated and downloaded from the server. This can be done by replacing one of them at a time, each requiring about 10kB. Visual words are continuously mounted to and unmounted from the leafs of the four kd-trees to adapt the partial vocabulary.

The retrieval and thus localization performance at the client is measured using the well known mean average precision (MAP) score, where all relevant panoramas have to be ranked first in the list of retrieval results to obtain a score of 1. In the following experiments we define all panoramas within a radius of 25 m around the query location to be relevant. Ground truth has been recorded using a state-of-the-art GPS receiver with manual post processing to correct for errors caused by multipath effects.

In Fig. 2 we evaluate the performance of different configurations of composing the partial vocabulary. As described in Section 3.1 K refers to the number of top ranked panoramas and radius R defines the maximum distance of panoramas whose visual words should be downloaded to the client. As neighboring panoramas overlap significantly with respect to the visual words, they are transmitted before sending the visual words of the next best ranked panorama. In this experiment we did not limit the data rate and thus all selected visual words are transmitted. The step size defines the number of frames until the server receives the next set of features from the client and updates the list of visual words to be downloaded. With increasing step size, the knowledge about the location of the client which has to localize itself during this time period based on the partial vocabulary decreases. Thus, when sending only the visual words that represent the feature set sent to the server (K0R0), the performance decreases already at the next frame. The best performance at almost all step sizes is achieved when sending the visual words of the best ranked retrieval result and the neighboring panoramas within a radius of 40 m (K1R40). Even at a step size of 1000 frames, where the client has to localize itself for 33 seconds based on one update from the server, we achieve a MAP of 0.53 which clearly outperforms the case when using a full vocabulary achieving a MAP of 0.34. A further increase of the partial vocabulary (K3R40) reduces the integration of prior knowledge and would ultimately lead to the performance of the full vocabulary. When including panoramas within a radius of 40 m the client can localize itself for about 55 m or about 1400 frames. After this point the performance drops for all configurations.

Assuming multiple clients sharing the bandwidth of a 3G network, we restrict the downlink data rate to 0.1, 0.25 and 1 Mbit/s in Fig. 3. At larger step sizes the client has sufficient time to continuously download the visual words and to integrate them into the retrieval system. However, for very short step sizes the rapid change of partial vocabularies can not be transmitted at the limited data rate. On average 9 panoramas can be found within a radius of 40 m resulting in about 6000 distinctive visual words. As shown in Fig. 4, this number of visual words only has to be transmitted if the location estimate at the server differs significantly from the last update. Thus, usually only the ids of the already sent visual words have to be transmitted. Further, with increasing time the number of visual

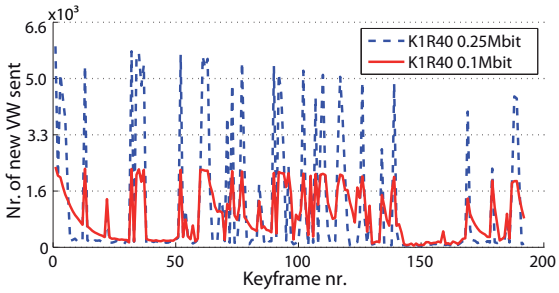

Fig. 4. Number of new visual words sent to the client per update (key frame) at two different data rates and a step size of 100 frames.

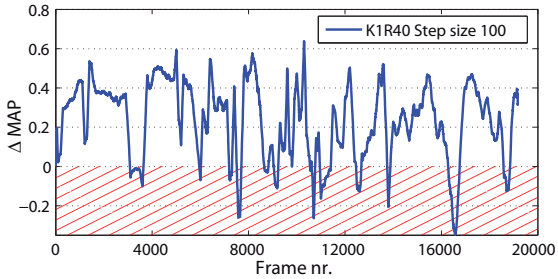

Fig. 5. Change in MAP for the K1R40 configuration at a data rate of 0.25 Mbit/s w.r.t. using the full vocabulary at individual frames.

words and associated inverted file entries increase which can be observed in the decay of the peaks in Fig. 4. Visual words that cannot be sent within the time of one update step might be required in the subsequent update and need to be transmitted as can be well observed when limiting the data rate to 0.1 Mbit/s.

To evaluate the performance of the approach over time we compute the MAP scores using a sliding window of 100 frames and compare them to the corresponding MAP scores of the full vocabulary in Fig. 5. For most frames a significant performance improvement can be observed. A reduced MAP score can occur when the retrieval results of the set of features uploaded to the server were poor and thus affect the partial vocabulary at the client.

When using the particle filter based composition of the partial vocabulary, as described in Section 3.2, a more stable and reliable estimate of the location can be computed at the server. This results in an additionally improved (overall MAP of 0.65) and less varying performance as shown in Fig. 6. Further, the number of visual words is significantly reduced to about 3000 as the area of location uncertainty is determined more precisely. This approach also facilitates very infrequent updates by the server with a step size of more than 700 frames, which results in an overall MAP of 0.62.

5. CONCLUSION

Preloading partial visual vocabularies onto the clients using the typically 5 times faster downlink of 3G networks allows us to perform mobile visual localization without communication delay close to real-time. Partial vocabularies are limited to the area of uncertainty about the location of the mobile device, which can be derived from features that are infrequently uploaded to the server. This not only reduces the amount of data to be downloaded but also significantly increases the retrieval performance as prior knowledge is integrated into the quantization process. Thus, the idea of partial vocabularies can also be used in other applications where prior knowledge should be integrated. We exploit database statistics to send

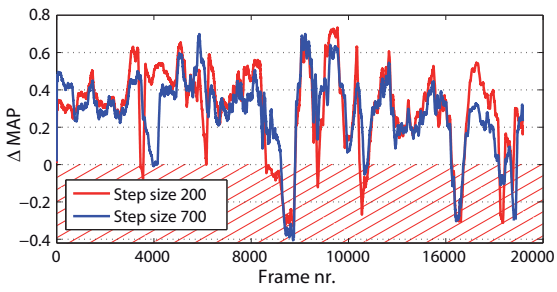

Fig. 6. Change in MAP using particle filter based partial vocabularies at 0.25 Mbit/s w.r.t using the full vocabulary at individual frames.

only those visual words to the client that provide most information about the location. Using a particle filter to fuse successive retrieval results at the server allows us to determine the area of uncertainty more precisely and thus increase the performance while reducing the amount of data to be downloaded.

6. ACKNOWLEDGEMENT

This research project has been supported by the space agency of the German Aerospace Center with funds from the Federal Ministry of Economics and Technology on the basis of a resolution of the German Bundestag under the reference 50NA1107.

7. REFERENCES

[1] "Google Street View," http://maps.google.com/streetview.

[2] "Microsoft Street-Side views," http://www.bing.com/maps/.

[3] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *IEEE Int. Conf. on Comp. Vision*, Nice, October 2003, pp. 1470–1477.

[4] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *IEEE Int. Conf. Comp. Vision Pattern Recognition*, Minneapolis, June 2007.

[5] B. Girod, V. Chandrasekhar, D. M. Chen, N. M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. S. Tsai, and R. Vedantham, "Mobile Visual Search," in *IEEE Signal Processing Magazine; Special Issue on Mobile Media Search*, July 2011, vol. 28, pp. 61–76.

[6] G. Schroth, R. Huitl, D. Chen, A. Al-Nuaimi, and E. Steinbach, "Mobile Visual Location Recognition," in *IEEE Signal Processing Magazine; Special Issue on Mobile Media Search*, July 2011, vol. 28, pp. 77–89.

[7] G. Schroth, S. Hilsenbeck, R. Huitl, F. Schweiger, and E. Steinbach, "Exploiting text-related features for content-based image retrieval," in *IEEE Int. Symposium on Multimedia*, Dana Point, December 2011.

[8] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod, "CHoG: Compressed histogram of gradients A low bit-rate feature descriptor," in *IEEE Int. Conf. on Comp. Vision and Pattern Recognition*, Miami, June 2009.

[9] D.M. Chen, S.S. Tsai, V. Chandrasekhar, G. Takacs, J. Singh, and B. Girod, "Tree histogram coding for mobile image matching," in *IEEE Data Compression Conference*, Snowbird, March 2009.

[10] Rongrong Ji, Ling-Yu Duan, Jie Chen, Hongxun Yao, and Wen Gao, "A lowbit rate vocabulary coding scheme for mobile landmark search," in *Intl. Conf. on Acoustics, Speech and Signal Proc.*, Prague, 2011.

[11] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *IEEE Int. Conf. Comp. Vision Pattern Recognition*, Minneapolis, June 2007.

[12] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, September 2004.

[13] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.