Poster Abstract: Efficient Visual Positioning with Adaptive Parameter Learning

Hongkai Wen, Sen Wang, Ronnie Clark, Savvas Papaioannou and Niki Trigoni

Department of ComputerScience, University of Oxford, Oxford, OX1 3QD, UK

***Abstract*—Positioning with vision sensors is gaining its popularity, since it is more accurate, and requires much less bootstrapping and training effort. However, one of the major limitations of the existing solutions is the expensive visual processing pipeline: on resource-constrained mobile devices, it could take up to tens of seconds to process one frame. To address this, we propose a novel learning algorithm, which adaptively discovers the place dependent parameters for visual processing, such as which parts of the scene are more informative, and what kind of visual elements one would expect, as it is employed more and more by the users in a particular setting. With such meta-information, our positioning system dynamically adjust its behaviour, to localise the users with minimum effort. Preliminary results show that the proposed algorithm can reduce the cost on visual processing significantly, and achieve sub-metre positioning accuracy.**

I. Introduction

Localisation with visual experiences in indoor environments has many advantages: it doesn't need infrastructure or accurate maps, and is more robust and accurate than other modalities such as WiFi. For instance, the Travi-Navi system [3] considers a teach-repeat navigation scheme, where a group of motivated users collects experiences of visual, radio, and magnetic measurements as they walk towards certain destinations, e.g. particular rooms. Later when another user tries to repeat, her position is estimated by comparing the live sensor observations with the stored experiences. This approach circumvents the difficulty of maintaining a globally consistent representation of the workspace, and thus is favourable in many application scenarios, such as personal guidance for the visually impaired, or remote assistance in industrial settings.

Unfortunately, the Travi-Navi system does not consider vision as the main positioning modality, but only samples sparse images for pathway identification. The major reason is that visual processing is prohibitively time-consuming for resource-constrained mobile devices. Fig. 1 shows the standard visual processing pipeline of the Bag-of-Words (BoW) image matching approach used by Travi-Navi. Given an observed image, the features are extracted, and further quantised into a vector of visual words with respect to a pre-trained vocabulary. Then the computed BoW vector is compared against a set of reference images (also in BoW format), where likelihood that two of them represent the same place is determined by certain distance metric. We find that on commodity mobile devices, feature detection and quantisation need a substantial amount of time to complete (>10s on Google Glasses and ∼5s on Nexus 6 phones), even with moderate resolution and a small vocabulary. On the other hand, to enable real-time positioning,

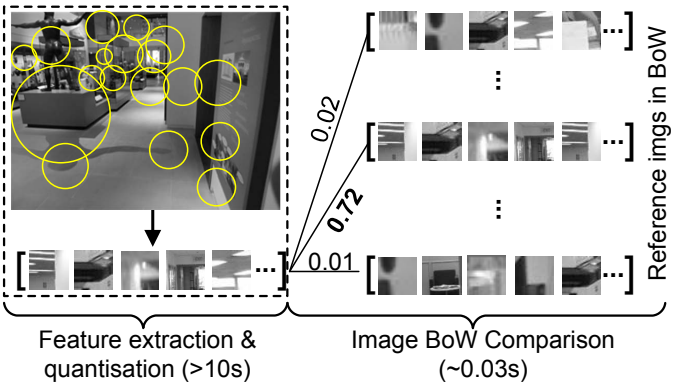

Fig. 1. Typical processing pipeline and running time of the Bag-of-Words image matching approach (estimated on Google Glasses with image resolution 800×600, using SURF [1] features and a visual vocabulary with ∼4000 words).

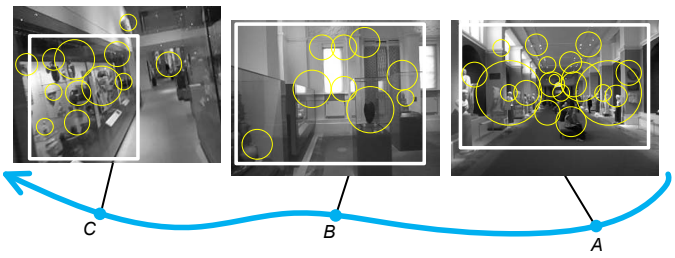

Fig. 2. Scene properties, e.g. feature distribution and type of visual elements, vary significantly across a visual experience.

one has to be able to process at least 1–2Hz, which is very challenging for the existing hardware.

To address this, we propose a novel adaptive parameter learning approach, which optimises the visual processing pipeline as more experiences are accumulated. The idea is that when following a previously taught route for multiple times, one should be able to discover clues on how to traverse it with much less effort. For example, Fig. 2 shows images sampled from three different places in a visual experience. Clearly, place A is large open space containing many features, while the scene at place B has only a few. This means at B, only a small subset of the visual vocabulary is sufficient for image matching. Also most features at B are close to the camera, and thus we can safely sample lower resolution images, which are much cheaper to process. In addition, at place C most features are clustered on the left (within the white box),

978-1-5090-0802-5/16/$31.00 ©2016 IEEE

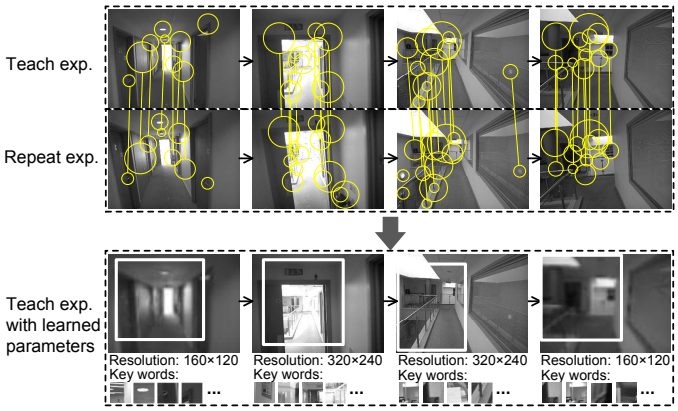

Fig. 3. An example of the proposed adaptive parameter learning approach.

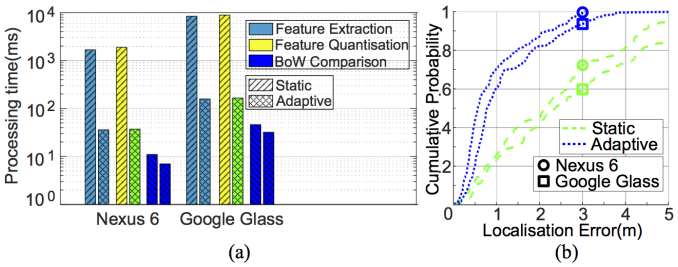

Fig. 4. (a) Breakdown of the visual processing time per image on different devices. The proposed algorithm with adaptive learning is up to 50× faster. (b) Error distribution of localisation on different devices.

and therefore it is possible to just process that small patch instead of the full image. Our approach follows this intuition, and learns how to spend the *minimum necessary* effort at different places to stay localised from repeating the previously taught experiences. The learned information is attached to the taught experiences, and is referred to when following those experiences in the future. In this way, we are able to massively reduce computation on visual processing, and make real-time visual positioning possible.

II. Proposed Approach

The proposed approach consists of two modules: a) a *localisation module* which runs locally at the mobile devices and positions the users with respect to the taught experiences, using the previously learned parameters; and b) a cloud side *learning module* which learns the optimal parameters for visual processing from the localisation results. The two modules essentially form a feedback loop: new parameters are learned through continued localisation of the users, which in return makes future positioning more efficient. Now we are in a position to explain the two modules in more detail.

Localisation with respect to taught experiences: At runtime, the metrical displacement of the user is computed based on data from the IMU sensors (i.e. accelerometer, gyroscope and magnetometer). The captured images are passed through the visual processing pipeline, and are matched against the images in the previously taught experiences. The matching results is then fused with the motion measurements through a state estimation algorithm, which determines the positions of the user with respect to the taught experiences. In the presence of any learned parameters, the visual processing pipeline is dynamically adjusted given the current state estimate. As the user moves, we also save the observed images and estimated displacement as a repeat experience.

Learning place dependent parameters: When localisation is finished, the saved repeat experience and the localisation results are uploaded to the cloud for parameter learning. Fig. 3 shows an example of how the learning process works. In our context, the results of positioning are the mapping between images of the repeat and teach experiences. For each image in the teach experience, our approach iteratively learns: a) a set of key visual words, which represent the most distinct scene elements; b) the minimum image resolution; and c) the salient region that contains the most informative features. Concretely, given the currently estimated resolution and salient region, the proposed learning algorithm computes the minimum set of key words, with which the known correspondence between teach and repeat experiences can still be maintained. The estimated key word set is then used to refine the salient region and image resolution, to further reduce the amount of pixels needed to be processed if possible. It alternates between these two processes until the parameters converge. Finally, the learned parameters are attached to the teach experience, which will be considered in next round of localisation.

III. Preliminary Results

We evaluated the proposed approach in two different indoor environments, an office building and a museum, with five participants of different genders, heights and ages. The participants wore Google glasses, and held mobile phones (Nexus 6) in their hands while walking. Ground truth is generated by applying map constrains to the collected inertial trajectory [2]. We compare the performance of the proposed approach with adaptive learning against the static approach which uses the same visual processing parameters throughout.

As shown in Fig. 4(a), by adaptively learning the optimal parameters for visual processing, the proposed approach is up to 50 times faster comparing to the static algorithm. This means we are able to process more images as the user moves, which could significantly improve the localisation accuracy, as shown in Fig. 4(b) (the mean localisation error is 0.96m). Therefore, with the proposed adaptive parameter learning approach, accurate positioning can be performed in real-time on resource-constrained platforms.

Acknowledgments The authors would like to acknowledge the support of EPSRC through grants EP/J012017/1 and EP/M019918/1.

References

[1] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Proc. ECCV*, 2006.

[2] Z. Xiao, H. Wen, A. Markham, and N. Trigoni. Lightweight map matching for indoor localisation using conditional random fields. In *Proc. IPSN*, 2014.

[3] Y. Zheng, G. Shen, L. Li, C. Zhao, M. Li, and F. Zhao. Travi-navi: Self-deployable indoor navigation system. In *Proc. MobiCom*, 2014.