Jakob Gaubatz*, Regine Hartwig, and Dirk Wilhelm

Location recognition in laparoscopic surgery

https://doi.org/10.1515/cdbme-2022-0030

Abstract: Navigation systems play an increasingly important role in minimally invasive surgery (MIS) by mitigating the problems rising from the decoupling of hand-eye movement of the surgeon. Many of these systems suffer from a high dependency on external optical tracking systems that require a constant line-of-sight to the optical markers being tracked. Simultaneous localization and mapping (SLAM) algorithms allow tracking the endoscope in cases where optical tracking fails due to the cluttered environment in the operating room. To ensure a correct camera pose estimate and to correct for drift, a recognition of previously visited locations (loop closures) is essential. We propose a method for location recognition in a minimally invasive scenario that only requires a stereo endoscope and an inertial measurement unit (IMU). We use a hierarchical bag-of-visual-words (BoW) algorithm that saves compact image representations and enables querying for matching images. A two-staged consistency check using a random sample consensus (RANSAC) and the data measured by the IMU ensure a high matching precision.

Keywords: location recognition, bag-of-visual-words, SLAM, laparoscopic surgery

1 Introduction

Minimally invasive surgery (MIS) has become a standard for many procedures. However, while patients benefit from decreased blood loss, pain, healing time, and improved cosmetic outcomes, the decoupling of hand-eye movement for navigational tasks outside of the surgeon's line-of-sight poses a big challenge for surgical staff and technology [1, 2]. Navigation systems try to mitigate these problems by enhancing visualization, planning, and tool tracking but face significant problems caused by the complex environment within the operating room. Most concepts rely on optical tracking systems using reflective spheres and an infrared (IR) tracking camera to estimate the camera pose and track surgical tools. Establishing the constant line-of-sight required for optical tracking is especially challenging in minimally invasive visceral surgery due to clutter, significant tool movements, and rotations. To tackle this problem, our algorithm updates the camera rotation and pose estimates if no IR tracking is available using a stereo endoscope and the data of an IMU attached to it (compare Fig. 1a) [3]. The visual feature-point-based simultaneous localization and mapping (SLAM) method builds a map of the surgical site and localizes the camera within it. Due to unavoidable measurement errors and noisy input this method suffers from a constant drift which decreases the camera pose estimate. A way to compensate for drift is the detection of loop closures (already visited locations). Loop closures are expected to occur regularly in MIS caused by large camera movements during exploration or while establishing optimal vision onto the surgical site. We built a database online from the images collected during the procedure and establish an efficient method to query for images that are similar to the currently observed image. Due to the large amount of image data produced during one procedure, we use a very compact image representation to enable a memory efficient storing of the images in a database.

This article presents a procedure to detect loop closures in an MIS scenery using a hierarchical BoW algorithm. We test the algorithm on a series of images of a visceral surgical scene and apply additional consistency checks using data of an IMU attached to the stereo endoscope.

2 Related Work

The efficient recognition of objects or locations in image data is a well-known problem in computer vision. Some location recognition algorithms use global features to generate a compact vector representing a scene. GIST being one of the most popular global descriptors, is used on several occasions for loop closure detection [4, 5]. It extracts information from images using Gabor filters at different frequencies and orientations. A different global image descriptor was proposed by Liu et al. [6]. They used an average of the U-V color space values of pixels enclosed in vertical edges to characterize an image. While this approach showed promising results in an indoor environment containing many vertical lines, it fails in a laparoscopic context, where straight lines are uncommon within the organic tissue structure.

In a BoW approach, other algorithms use local feature descriptors like SIFT, SURF, or FAST to create a compact image representation. Originating from text-based document analysis, BoW uses a dictionary of visual words generated by clustered local image features. Counting the number of occur-

***Corresponding author: Jakob Gaubatz,** Research Group MITI, Technical University of Munich, Munich, Germany, e-mail: jakob.gaubatz@tum.de
Regine Hartwig, Dirk Wilhelm, Research Group MITI, Technical University of Munich, Munich, Germany

rences of each visual word (cluster center) in an image enables a compact image representation. This approach has become one of the gold standards in location recognition [7–9]. To increase efficiency for large-scale object recognition, Nistér et al. [10] suggested a tree-shaped vocabulary that enables a faster lookup of visual words and the efficient use of a more extensive, more discriminative vocabulary. Gálvez-López et al. [9] improved this approach by the use of binary BRIEF descriptors with FAST keypoints to speed up the computational bottleneck of image descriptor extraction. In a medical context, Moll et al. [11] performed feature matching using a BoW approach to re-initialize their algorithm for rotating laparoscopic images.

To our knowledge, our proposed method is the first BoW procedure to detect loop closures, specially designed for a laparoscopic context. Our contribution is as follows: Firstly, we built a vocabulary from laparoscopic images since feature descriptors differ from publicly available datasets of human-made indoor or outdoor environments. Similarly to [9] we decided to use a hierarchical BoW dictionary that discretizes a binary descriptor space but we follow the approach of [12] and use ORB-descriptors to overcome the described problems of lack of rotation and scale invariance when using the FAST+BRIEF features. Secondly, we evaluated appropriate vocabulary size and thresholding parameters. Thirdly, in contrast to a temporal consistency check proposed in [9], we use the measured angular velocity data of the IMU attached to the endoscope.

3 Method

3.1 Vocabulary Tree

To extract the BoW dictionary, the descriptor space gets divided into N visual words. This visual vocabulary enables the formation of a compact image representation by assigning all image descriptors to the closest word. The number of occurrences of each word in an image results in a sparse histogram called BoW vector $v_i \in \mathbb{R}^N$ that represents the image. The vector of each frame is saved and enables querying for similar images without the need to compare image descriptors. We chose a binary descriptor space using ORB-descriptors for speeding up descriptor computation. The tree-shaped vocabulary with leaf nodes as the visual words reduces the number of computations required to map an image into the BoW space by traversing the tree from the root to the leaves. At each layer, we select the node which minimizes the Hamming distance.

We created a specially designed vocabulary offline, performing k-means clustering on a set of training image descriptors. The resulting clusters build the first layer of the tree. We get subsequent levels by repeating the procedure to obtain L levels and $N = k^L$ visual words. Each word gets assigned a weight based on its occurrence frequency within the dataset, giving a higher weight to less frequent and thus more discriminative words. We use the popular *term frequency-inverse document frequency* as suggested by [13].

3.2 Thresholding

To compare the similarity between two BoW vectors v_1 and v_2 we calculate the L_1-distance:

$$s(v_1, v_2) = \left\| \frac{v_1}{|v_1|} - \frac{v_2}{|v_2|} \right\|_1 \tag{1}$$

If the resulting score is beneath a certain threshold x_{th}, the two images should show a similar content, thus representing a loop closure. A sophisticated choice of the threshold is crucial to ensure the best possible performance of the BoW algorithm. In our binary classification task a suitable threshold can be found by analyzing the probability density functions (PDF) $f_0(x)$ (images represent a loop closure/ground truth positive) and $f_1(x)$ (images represent no loop closure/ground truth negatives) for a given dataset. When choosing a threshold x_{th} the *true positive rate* (TPR) is given by 2 and the *false positive rate* (FPR) is given by 3:

$$TPR(x_{th}) = \int_{-\infty}^{x_{th}} f_0(x)dx \tag{2}$$

$$FPR(x_{th}) = \int_{-\infty}^{x_{th}} f_1(x)dx \tag{3}$$

We choose the threshold by the maximum likelihood decision rule, resulting in an x_{th} where $f_0(x)$ = $f_1(x)$.

Additionally, a *receiver operating characteristic curve (ROC-curve)* can be generated by plotting TPR over FPR for different thresholds x_{th} that helps to determine a trade-off between FPR and TPR by choosing an error level, i.e, an upper bound for the false positive rate.

3.3 Consistency Checks

For each image pair (I_a, I_b) obtained from the BoW query, with $s(v_a, v_b) < x_{th}$ we apply additional consistency checks to confirm a valid loop closure. First, we get a camera pose by matching at least six 2D-3D feature correspondences between both images. For each landmark x, we find a 3D-location $P_{a,x} \in \mathbb{R}^3$ w.r.t. the camera location of the initialization image I_a. The required 3D-locations are obtained using stereo-correspondences found by descriptor matching and

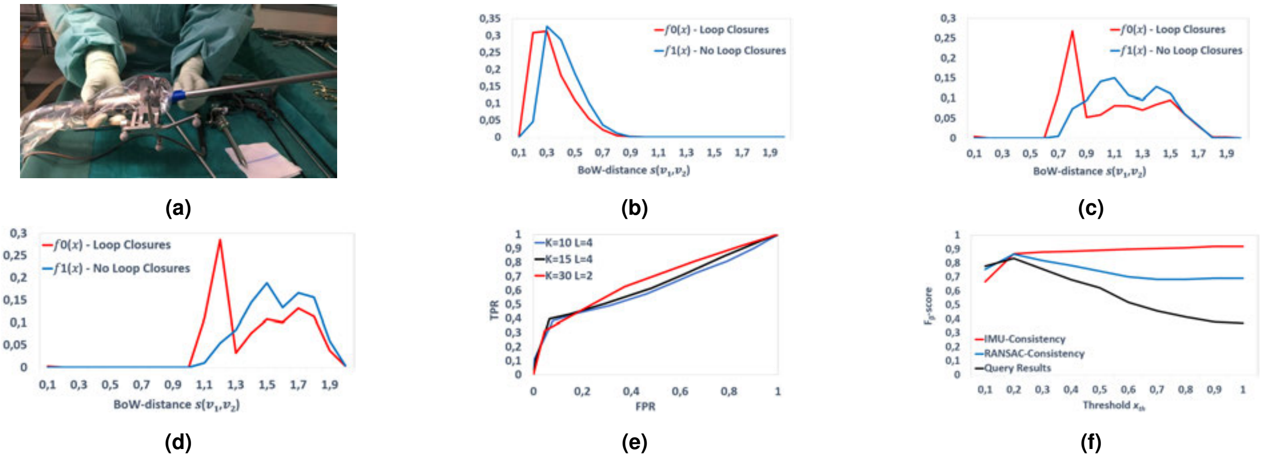

Fig. 1: (a) The used setup: A stereo endoscope with attached IMU and IR sensor. (b) PDF, calculated based on the BoW-distances between all images in the test dataset for dictionary parameters $k = 30$, $L = 2$, $N = 30^2 = 900$, (c) $k = 10$, $L = 4$, $N = 10^4 = 10000$ and (d) $k = 15$, $L = 4$, $N = 15^4 = 50625$. (e) *ROC-curves* derived by TPR and FPR for different matching thresholds x_{th} and dictionary parameters. (f) F_β-scores after match querying (black), RANSAC consistency check (blue) and IMU consistency check (red).

reconstruction of the depth using the known camera calibration of our stereo endoscope. 2D feature points $p_{b,1}, ..., p_{b,n}$ in image I_b are added as corresponding to $P_{a,x}$ if the descriptor distance is smaller than a threshold and the distance gap to all other landmarks is larger than a threshold. The camera pose can then be calculated using a perspective-n-point algorithm (PnP) [14] inside a RANSAC provided by OpenGV.

If we find a suiting camera pose, we compare the estimated camera rotation $R_{\text{RANSAC}} \in \text{SO}(3)$ to the integrated gyroscope measurements $R_{\text{Gyro}} \in \text{SO}(3)$ attached to the endoscope to check if the real-world data supports the RANSAC output. The distance between two Lie-Group elements in $\text{SO}(3)$ represented as the norm of an element of the Lie algebra $\mathfrak{so}(3) \cong \mathbb{R}^3$ is given by:

$$d(R_{\text{RANSAC}}, R_{\text{Gyro}}) = \| \log \left({R_{\text{RANSAC}}}^{-1} R_{\text{Gyro}} \right) \| \in \mathbb{R} \quad (4)$$

If $d(R_{\text{RANSAC}}, R_{\text{Gyro}})$ lies under the threshold $d_{th} = 0.5rad$, I_a and I_b are considered to represent a loop closure.

4 Evaluation

To evaluate the performance of our algorithm, we extracted three dictionaries with different parameters L and k from a set of images of the MITI dataset [15]. The dataset contains images from inside the abdomen acquired during a visceral MIS. A set of 39233555 ORB-descriptors extracted from a series of 75702 images using the OpenCV library form the training base for the dictionaries. We perform a hierarchical k-means clustering using the DBoW3 library [16]. To compare the performance of the extracted dictionaries and to choose a suitable similarity threshold x_{th}, we used a series of 828 images containing a total of eight loop closing scenarios, priorly excluded from the training dataset. The loop closures show different anatomical structures and exhibit tissue deformations and illumination changes between the closing events, as it is expectable during an intervention. Fig. 1b, 1c, and 1d show the probability density functions of the dictionaries calculated based on the test dataset. The curve's separability differs depending on the number of parameters of the BoW dictionary. For an increasing number of parameters we observe that the vocabulary fails to meaningfully represent the descriptor space for unseen data. The corresponding *ROC-curves* in Fig. 1e show an improved generalizability of the vocabulary with fewer parameters on the test dataset. We obtain the best trade-off between FPR and TPR using the smallest dictionary with $k = 30$ and $L = 2$. Based on the maximum likelihood decision rule, we choose the line intersection threshold $x_{th} = 0.3$ in Fig. 1b which results in a $TPR = 0.63$ and $FPR = 0.37$ and corresponds to the shoulder point in the *ROC-curve*. Alternatively, we can choose a predefined error level and design a Newman-Pearson test that decides for the hypothesis of a match being valid based on a threshold determined by the chosen FPR.

Since the BoW query output is used in a SLAM algorithm, the number of false detections (false positives) should be as small as possible to give a high rating to the query results when weighting the SLAM-residuals. Therefore, we attach more importance to high precision than to high recall using the F_β-score:

$$F_\beta = (1 + \beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall} \quad (5)$$

The parameter β adjusts the weighting of recall and precision. A $\beta < 1$ increases precision weighting, and a $\beta > 1$

increases the weighting for recall. In our case, we choose $\beta = 0.1$. We evaluate the metric on the same set of 828 images previously used. We define a true positive as both frames being part of a matching frame interval also considering neighboring frames. The matching frame intervals were hand-labeled prior to testing. Like in a real-world scenario, we process all image frames successively by first converting them into the BoW space, saving the BoW vector in the database, and then querying for matches. Since not all images are present in the database from the beginning, the number of ground truth matches for one query frame is the number of matching frames present in the BoW database at query time.

Fig. 1f shows the F_β-scores after querying for a match, after the consistency check using RANSAC, and after the last consistency check with the angular velocity evaluated for different matching thresholds x_{th}. It is observable that a considerably high score can be achieved without consistency checks when using a very restrictive threshold. Since a low threshold also causes a high rate of discarded loop closures, taking a low threshold is not optimal, proving the necessity of consistency checks. The RANSAC consistency check allows the choice of a much higher threshold without the drastic decrease in performance. The PnP, however, holds the drawback that for significant tissue deformations in the image sequence it can not find a rigid transformation between the matched frames, thus decreasing the recall. To compensate for that, we choose the RANSAC and PnP hyperparameters to be less restrictive, increasing the recall but simultaneously decreasing precision since it is more likely to find meaningless transformations. We filter out those false positives by the IMU-consistency check that constantly keeps high F_β-scores even for increasing thresholds x_{th}. The recall decreases in cases of large, rigid object movements, where descriptors are matching but IMU measurements do not fit the observed movement.

5 Conclusion

We present a method for location recognition to detect loop closures in laparoscopic MIS using a stereo endoscope and an IMU attached to it. Using a specially designed BoW vocabulary and approach, we can effectively store compact image representations and query for similar images. The evaluation demonstrates that the two-staged consistency check using RANSAC and the angular velocity data measured by the IMU provides high matching precision even for less restrictive matching thresholds. However, the main limitation of our method is a low recall for significant tissue deformations or large object movements. Therefore, the consistency check serves as a segmentation of rigid image parts, downweighting/deleting residuals of matches from dynamic/deformed objects from the SLAM optimization problem.

In the future our method will be integrated into an already existing SLAM system [3] to further validate the impact onto the camera tracking accuracy.

References

[1] Driessen SRC, Sandberg EM, Rodrigues SP, van Zwet EW, Jansen FW. Identification of risk factors in minimally invasive surgery: a prospective multicenter study. Surg Endosc. 2017;31(6):2467-2473.

[2] Rodrigues SP, Wever AM, Dankelman J, Jansen FW. Risk factors in patient safety: minimally invasive surgery versus conventional surgery. Surg Endosc. 2012;26:350-356.

[3] Hartwig R, Ostler D, Rosenthal JC, Feußner H, Wilhelm D, Wollherr D. Constrained Visual-Inertial Localization With Application And Benchmark in Laparoscopic Surgery. arXiv - CS - Computer Vision and Pattern Recognition (IF). 2022.

[4] Singh G, Košecká J. Visual Loop Closing using Gist Descriptors in Manhattan World. ICRA Omnidirectional Vision Workshop, Anchorage. 2010.

[5] Liu Y, Zhang H. Visual loop closure detection with a compact image descriptor. 2012 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems. 2012;1051-1056.

[6] Liu M, Siegwart R. Topological Mapping and Scene Recognition With Lightweight Color Descriptors for an Omnidirectional Camera. IEEE Trans. on Robotics. 2014;30:310-324.

[7] Angeli A, Filliat D, Doncieux S, Meyer J. Fast and Incremental Method for Loop-Closure Detection Using Bags of Visual Words. IEEE Trans. on Robotics. 2008;24:1027-1037.

[8] Cummins MJ, Newman P. Appearance-only SLAM at large scale with FAB-MAP 2.0. The Int. Journal of Robotics Research. 2011;30:1100-1123.

[9] Gálvez-López D, Tardós JD. Bags of Binary Words for Fast Place Recognition in Image Sequences. IEEE Trans. on Robotics. 2012;28:1188-1197.

[10] Nistér D, Stewénius H. Scalable Recognition with a Vocabulary Tree. 2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. 2006;2:2161-2168.

[11] Moll M, Koninckx TP, Van Gool LV, Koninckx PR. Unrotating images in laparoscopy with an application for 30° laparoscopes. 2009.

[12] Mur-Artal R, Tardós JD. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Transactions on Robotics. 2017;33:1255-1262.

[13] Sivic J, Zisserman A. Video Google: a text retrieval approach to object matching in videos. Proc. IEEE Int. Conf. on Computer Vision. 2003;2:1470-1477.

[14] Lepetit V, Moreno-Noguer F, Fua P. EPnP: An accurate O(n) solution to the PnP problem. Int. Journal of Computer Vision. 2009;81(2):155–166.

[15] Hartwig R, Ostler D, Rosenthal JC, Feussner H, Wilhelm D, Wollherr D. Miti: Slam benchmark for laparoscopic surgery. TUM. 2021.

[16] DBoW3 [computer software]. 2017. Available: https://github.com/rmsalinas/DBow3