POLITECNICO MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

NeuroGlue: Attentional Graph Based Mosaicking in Neurosurgery

LAUREA MAGISTRALE IN BIOMEDICAL ENGINEERING - INGEGNERIA BIOMEDICA

Author: ANNA MARIA DE LUCA

Advisor: PROF. ELENA DE MOMI

Co-advisor: ALESSANDRO CASELLA

Academic year: 2021-2022

1. Introduction

The Surgical Microscope (SM) is the gold standard instrument in neurosurgery. The SM allows for visualization of the surgical field and the anatomical details of the brain structures. On the other hand, the high magnifications provided by the SM cause a very limited field of view (FoV) that may lead to harmful operations on anatomical structures or a nearby organ, which will affect the surgical outcome. [3]
In neurosurgery, most of the operations involve tumor resection or brain lesions removal. Intraoperative navigation systems are needed to detect brain tumor or lesion position based on its coordinates in the preoperative image. However the main problem of this instrument is the error introduced by the brain shift, a nonrigid brain transformation that occurs after craniotomy. [3]
In the opinion of Humanitas Hospital neurosurgeons, the illustrated limitations have an evident influence on Cerebral Cavernous Malformations (CCMs) removal and glioma resection.
A reduced FoV could actually prevent the surgeon from viewing the brain tumor or the lesion entirely, and navigation system errors may affect their localization. These are relevant factors that make neurosurgical procedures still challenging.

An intra-operative system able to perform a real-time and broad exposure of the surgical theatre could be an effective tool to support neuro-surgeons. Computer-Assisted Intervention (CAI) is a powerful ally to deal with these types of challenges through new developing techniques like Artificial Intelligence (AI) and Deep Learning (DL).
In particular, mosaicking can extend the limited neurosurgical FoV by creating a panorama of the surgical environment.
This expansion of the operative field could be practiced at any stage of the surgical intervention generating a reconstruction of the brain's superficial layers. Mosaicking is achieved by performing microscope video frame registration without introducing any external sensor in the operating room. The panorama is computed in real-time, and it can represent an important reference for the surgeon during the procedure. Indeed the surgeon can work on the brain, observing each anatomical detail thanks to the high magnifications of the SM and, at the same time, have a broader view of the entire scene without further SM movements.
This tool could be integrated effortlessly into surgical workflow thanks to its ease of use and robustness.
The contribution of this work can be summa-

rized as follows:

1. To the best of our knowledge, it represents the first application of mosaicking on a neurosurgical dataset.
2. A robust self-supervised method for keypoints detection and description.
3. An Attentional Graph Neural Network (AGNN) trained in a self-supervised way on intra-operative images for keypoints matching.
4. A homography filter to avoid reconstruction errors due to unexpected events.

2. Materials and Methods

The purpose of video mosaicking is to combine consecutive frames of a video sequence, in which each frame shows only a partial local view of the field of interest. It allows obtaining a broader view of the same scene.

The following four stages characterize the classical mosaicking approach: i) Keypoints detection and description; ii) Keypoints matching with outlier rejection; iii) Homography estimation; and iv) Image warping and blending. In this process, each step is essential for the correct execution of the next ones.

A features detector is an algorithm that searches for keypoints in an image. Keypoints can range from a single pixel to edges, corners, contours, blobs, junctions, and lines; they are expressed by a system of coordinates and represent the most significant pixels of the selected image.

Once the keypoints have been extracted from the image, the feature description phase is applied. A descriptor is necessary to assign a distinctive identity to each key point, which allows their effective recognition for matching. The features description is based on unique patterns possessed by the neighboring pixels of each keypoint.

Given a frame sequence, each image pair is considered. The frame pair consists of a moving image (B) registered with respect to a fixed image (A).

Keypoints detection and description are computed for each image of the pair. The next stage of the mosaicking process is the feature matching between the keypoints of the two images. It aims at establishing correct correspondences from the keypoints sets.

Afterwards, RANdom Sample Consensus (RANSAC) algorithm is employed for the homography estimation. The homography matrix is applied to the moving image B and is used to merge A with B. This step is called image warping. [5]

After this overview of the overall mosaicking process, the proposed NeuroGlue architecture is described in the following sections.

2.1. Keypoints Detection and Description

In NeuroGlue, the keypoints detection and descriptor phases are combined in a Fully Convolutional Neural Network (FCNN), which is called SuperPoint. [1]

In particular, this network is able to detect robust and repeatable keypoints and attach a fixed dimensional descriptor vector to each keypoint for further processing, such as image matching.

The first step that is applied is the dimensionality reduction of each input image, which is performed with a single shared encoder. After the encoder, the architecture splits into two decoders: one for keypoint detection and the other for keypoint description. Both decoders operate on a shared and spatially reduced representation of the input performed in a VGG-style.

In this way, keypoints and descriptors are jointly learned thanks to the shared encoder. Indeed this represents an improvement compared to the classical methods, which first detects keypoints, then computes descriptors lacking the ability to share computation and representation across the two tasks. [1]

For the image pair considered, the keypoints p^A and the relative descriptors d^A for the frame A and p^B with the descriptors d^B for the frame B are obtained (Fig. 1).

2.2. Graph Based Matching

The keypoints matching is performed through an Attentional Graph Neural Network architecture. In the first step, keypoints and descriptors, obtained from the SuperPoint network, are subsequently combined into a single vector using a keypoints encoder. In this way, a first features representation is achieved.

Afterwards, Attentional Aggregation is computed. In neural networks, attention is a technique that mimics cognitive mechanisms. It is based on an encoder-decoder architecture which

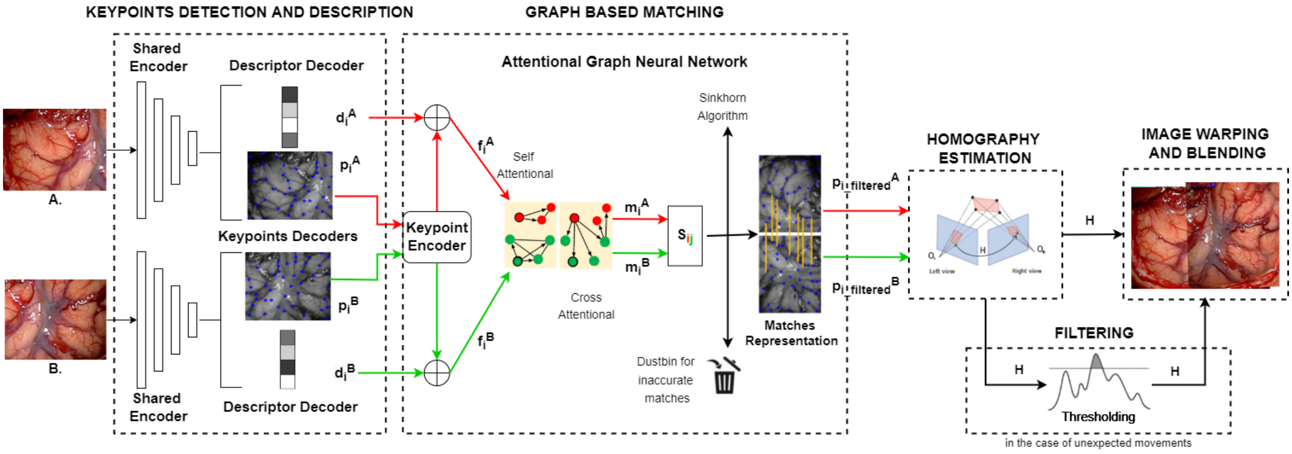

Figure 1: Overview of the proposed framework for neurosurgery mosaicking, as described in Sec. 2. The first block concerns keypoints detection and description phases. Each keypoint and descriptor of the image A are indicated with p_i^A and d_i^A respectively. The same idea is applied to the image B (Sec. 2.1). These outputs are then combined ($\bigoplus$) using a keypoint encoder in order to obtain a unique feature vector for each image (f_i^A and f_i^B). m_i^A and m_i^B are the matching descriptors obtained from the alternation of Self and Cross Attentional Layers. The affinity between the correspondences is represented by the score matrix S_{ij}, which is also used to filter out invalid matches with a dustbin. Matching optimization is performed with the Sinkhorn Algorithm. (Sec. 2.2) Removing the key points related to invalid matches, $p^A_{i_{filtered}}$ and $p^B_{i_{filtered}}$ are employed for the homography estimation (H)(Sec. 2.3), essential for image warping and blending (Sec. 2.4). Also the filtering stage is represented and it is described in Sec 2.5.

identifies the most relevant points into an input data, by assigning them a higher weight factor respect to the less important ones. In particular, there is the Self Attentional Aggregation in which each keypoint is related to another keypoint of the same image. This technique allows to focus on a subset of keypoints basing on their locations. Instead, with the Cross Attentional Aggregation, each keypoint is related to a keypoint belonging to the other image of the considered pair (Fig. 1) in order to generate potential matches. The alternation of the two Attentional Aggregation layers allows to develop the keypoints connection, making it stronger and more stable for matching. From this process, matching descriptors (indicated with m_i^A and m_i^B) are achieved.

The affinity between the correspondences is defined with a score matrix (S_{ij}), which is computed by the dot product between the matching descriptors of each image, as Fig. 1 indicates. The score matrix is also used to filter out the invalid matches and the relative keypoints, present due to occlusions and noise. This procedure is carried out with the introduction of a dustbin, represented in Fig. 1.

Finally, the Sinkhorn algorithm is applied as an optimization layer in order to increase the reliability of the optimal transport estimation for matching computation. Indeed optimal transport tool is used to find the minimal cost in probability distribution data pairs. Sinkhorn algorithm is an iterative process that normalizes the score matrix around the rows and the columns. From the obtained result, the matches can be extracted.[4]

2.3. Homography Estimation

After that the valid matches and the correspondent keypoints ($p^A_{i_{filtered}}$ and $p^B_{i_{filtered}}$) are extracted, RANdom Sample Consensus (RANSAC) and Levenberg-Marquardt (LM) algorithms are jointly applied to estimate the homography. The objective of RANSAC is to select the optimal set of keypoints and filter out outliers that do not fit with a defined type of transformation.

The homography H is a 3×3 matrix, which provides the relative transformation of B with respect to A. In general, the homography matrix represents the transformation between the points of two different planes. [5] The homography is used to project the 3D movements of the camera into a 2D space. Indeed it is created combining different components: the camera intrinsic matrix (which depends on the focal length); rotation matrices around the X,Y,Z axis and the translation array in X,Y,Z directions.

2.4. Image Warping and Blending

Image warping and blending are performed by combining B with A, basing on the matrix H, previously computed, until all matched feature points are aligned. In this way, a partial panorama is obtained, as the Fig. 1 shows. [5]

2.5. Filtering

Neurosurgery procedures are characterized by limited movements of the microscope. Indeed broad and rapid movements are unlikely because surgeons used to work with high magnifications in a reduced operative field. However, it could happen to mistakenly impact the microscope, generating very fast movements that could make the registration algorithm fail. Any abnormal movement of the camera (either translation, rotation or scaling) could generate distortions and reconstruction errors. For this reason, a filtering stage is implemented.

After H estimation, the Singular Value Decomposition (SVD) is performed. SVD procedure consists of the matrix factorization using eigenvalues and eigenvectors. Experimentally it was demonstrated that if an unexpected movement occurs, one or more decomposed values show a steep increase characterizing an abnormal homography.

In particular with SVD, six parameters are obtained: t_x and t_y which reflect the translating movements of the camera, s_x and s_y, related to the scaling and γ and θ that translate the rotational transformations. The correlation among the different parameters is assessed by computing the Pearson's Correlation (ρ) reported in Table 1.

Two parameters are correlated when the value of ρ is close to ±1. For this reason γ is selected. To achieve a more complete and robust analysis also t_x is considered.

At each iteration, t_x and γ are compared with two thresholds, respectively, which are experimentally selected. An abnormal change in the homography (one or both values are over-threshold) interrupts the registration procedure, discarding the associated frame since a new valid frame is present. Therefore this homography check allows for the mosaic recovery, obtained before the unexpected event occurs.

Table 1: Pearson correlation (ρ) among parameters obtained through SVD of the homography transformation computed in Sec. 2.5

ρ	t_x	t_y	s_x	s_y	γ	θ
t_x	1	0.999	0.999	-0.999	0.182	0.997
t_y	0.999	1	0.999	-0.999	0.169	0.996
s_x	0.999	0.999	1	-0.999	0.190	0.996
s_y	-0.999	-0.999	-0.999	1	-0.151	-0.995
γ	0.182	0.169	0.190	-0.151	1	0.223
θ	0.997	0.996	0.996	-0.995	0.223	1

2.6. Dataset

The dataset available was provided by Humanitas Research Hospital of Milano, in which several videos were captured from a Carl Zeiss Surgical GmbH microscope. In particular, three videos (called Video1, Video2, Video3) are employed for the NeuroGlue validation. In particular Video1 contains 1677 frames, Video2 is composed by 667 frames (it is the shorter one) and Video3 has 3543 frames. The extracted frames have original dimension of 720 × 576. Images captured from a surgical setting are characterized by regular patterns (for the blood vessel's structure), viewpoint changes, illumination variations and motion blur, factors that make the classical mosaicking method not robust and stable enough.

3. Experimental Protocol

3.1. Training Phase

The NeuroGlue training is performed on a dataset of 6144 non-overlapped patches with dimension of 256 × 256 pixels.

NeuroGlue is trained end-to-end in self-supervised way [4] for 300 epochs.

The keypoints detection and description network training is based on a method called Homography Adaptation. [1] It consists on the random homographies generation that are used to warp copies of the input image and combine the results. The same image is deformed L times, using L different random homographies. In each warped image the keypoints are extracted, then

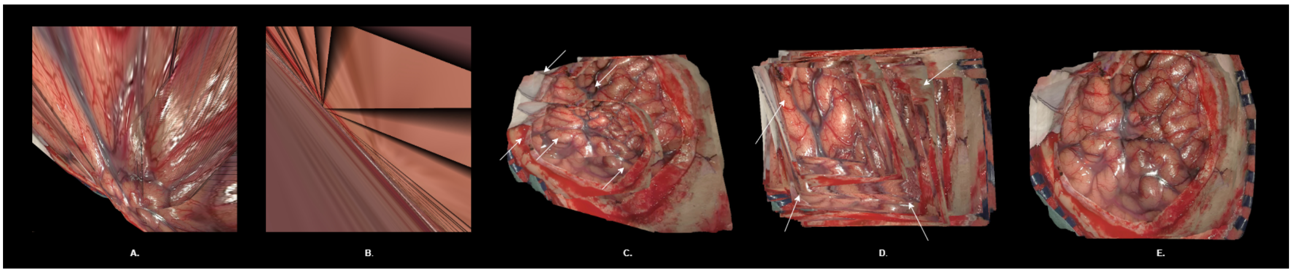

Figure 2: The figure shows the mosaics computed using A) BRISK, B) ORB, C) SIFT, D) SuperPoint, E) NeuroGlue. The white arrows indicate some inaccuracies and alignment errors in the registration. In this figure, the results of Video3 are reported.

the images are unwarped, so an inverse transformation is applied to restore the input image and finally all the L information obtained are merged to create the keypoints set.

The Attentional Graph Neural Network training is developed by carrying out homography matrix computation and random image warping. In particular, considering one patch X randomly selected, its keypoints and descriptors are computed with the previously trained SuperPoint architecture. [1]

A random warp transformation is applied to X: the image is deformed taking into account the expected movements of the microscope camera. In this way X_{warped} is obtained.

Afterwards, having X and X_{warped}, the homography matrix M between the two frames is computed. This is used to map the keypoints previously calculated with SuperPoint in X_{warped}.

In this way the network learns to generate correct matches, since the keypoints correspondence is specially guaranteed, using the homography matrix M.

3.2. Ablation Study

To the best of our knowledge this work represents the first mosaicking application on a neurosurgical environment. For this reason, the analysis concerns the quality with which the methods presented in literature for keypoints detection and matching (both traditional and learning based) fit a neurosurgical dataset.

The proposed method is compared with three traditional features detectors and descriptors: Binary Robust Invariant Scalable Keypoints (BRISK), Oriented FAST and Rotated BRIEF (ORB) and Scale Invariant Feature Transform (SIFT). These keypoints detectors are coupled with the K-Nearest Neighbors (kNN) matching algorithm. [5]

It was tested also the SuperPoint network for keypoints detection, pre-trained with the COCO dataset [1] coupled with KNN algorithm for matching. SuperPoint architecture represents one of the most promising and accurate technique in the state of the art for learning-based keypoints detection. [2].

The compared methods can be indicated as follows:

- Method 1: BRISK + KNN
- Method 2: ORB + KNN
- Method 3: SIFT + KNN
- Method 4: SuperPoint + KNN
- Method 5: NeuroGlue

3.3. Evaluation Metric

For the evaluation of the panorama reconstruction, obtained with the different methods, 5-frames Structural Similarity Measure (i.e. *SSIM*, indicated as s in Equation 1) is computed. From the frame sequence of each video, one frame every five is selected. From this sampled list, consecutive pairs are taken. The second frame is transformed according to the relative transformation with the first and the metric is computed basing on the following equation:

$$s_{i\to i+n} = \text{sim}\left(w(\tilde{I}i, Hi \to i+n), \tilde{I}_{i+n}\right) \quad (1)$$

where n is equal to 5, w is the warping function, sim is a similarity function, I_i is the first image, I_{i+n} is the second image and H is the relative transformation.

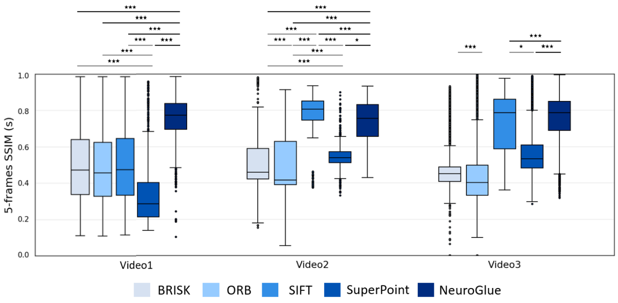

Figure 3: Boxplot of 5-frames *SSIM* (s) are represented for the tested methods: BRISK (in light grey), ORB (in light blue), SIFT (in blue), SuperPoint (in dark blue) and NeuroGlue (in night blue). 5-frames *SSIM* is computed as described in Sec. 3. Wilcoxon Signed-Rank test outcomes are present. The results of Video1, Video2 and Video3 are reported.

4. Results and Discussion

A visual comparison of the mosaics of Video3 obtained with the discussed methods is reported in Fig. 2. It is possible to observe that the traditional mosaics of Video3 show several alignment errors and spatial distortions. SIFT presents a better result respect to ORB and BRISK, in which the mosaics are completely hidden by the deformations, because it is characterized by a better scale and rotation invariance [5].

These inaccuracies are reflected in the boxplots of Fig. 3, where *SSIM* of BRISK, ORB and SIFT results on average very low, specially if compared with the NeuroGlue *SSIM* values.

In general the classical feature detectors may fail to extract enough keypoints into images which are characterized by repetitive patterns, illumination variation, and motion blur, typical conditions of a neurosurgical dataset. They are primarily employed for reconstruction of landscapes, buildings or everyday life objects.

The mosaic obtained with SuperPoint shows several errors in Fig. 2 and this provides lower values of *SSIM* as reported in Fig. 3. Despite the promising keypoints detection network, the peculiar characteristics of a neurosurgical images, as previously stated, and the lack of perfect coupling between the detection and matching phases lead to inaccurate results.

NeuroGlue panorama stands out clearly from the other methods, this is due to the stability and the strength of the network (demonstrated with the lower variance of *SSIM* Fig. 3), which learned to deal with neurosurgical images, thanks to the adapted training. The same discussion can be applied to the other two videos, observing the boxplots in Fig. 3. The only difference is present in Video2, in which SIFT results quite accurate. The reason is that the video is significantly shorter than the others, as explained in Sec. 2.6. Using longer videos so passing more times up to the same areas (like is done in Video1 and Video3), demonstrates the instability of SIFT respect to the proposed method. The Wilcoxon Signed-Rank test is performed, as Fig. 3 indicates, and it underlines a significant difference between NeuroGlue and the other methods.

5. Conclusion

The peculiar conditions of a surgical setting could affect the mosaicking quality. This issue introduces the need to find stable keypoints detectors and establish stronger connections between the keypoints of consecutive frames.

The proposed method showed to achieve better performance in terms of *SSIM* compared to the traditional feature detection algorithms and also respect to the SuperPoint method, underling the importance of the domain adaptation.

The introduction of an intra-operative system for real-time FoV expansion could represent a valuable tool to deal with low visibility issue in neurosurgery. The obtained panorama can represent an important reference for surgeon as it allows him to observe the brain tissue details using the SM magnifications and at the same time, to consult a broader map of the operating field, without moving the SM.

It was tested to be applied in the early stage of surgery to generate a reconstruction of the brain superficial layers. However the promising results show that this method could also be employed for the reconstruction of deeper anatomical structures.

Testing the proposed framework in a broader dataset could be useful to validate even more the network and the system integration with intra-operative navigators or pre-operative MRI could be beneficial for lesions or tumour localization during surgery.

References

[1] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. 2018.

[2] Cuiyin Liu, Jishang Xu, and Feng Wang. A review of keypoints' detection and feature description in image registration. 2021.

[3] Ling Ma and Baowei Fei. Comprehensive review of surgical microscopes: technology development and medical applications. 2021.

[4] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. 2020.

[5] Shaharyar Ahmed Khan Tareen and Zahra Saleem. A comparative analysis of sift, surf, kaze, akaze, orb, and brisk. 2018.