Advances in Computer, Signals and Systems (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2024.080311
ISSN 2371-8838 Vol. 8 Num. 3

Attention-based mechanism for SuperPoint feature point extraction in endoscopy

Mingyue Zhang[1,a,*]

[1]*School of Information and Electronic Technology, Key Laboratory of Autonomous Intelligence and Information Processing in Heilongjiang Province, Jiamusi University, Jiamusi, China*
[a]*1786969470@qq.com*
[*]*Corresponding author*

Keywords: Attention mechanism, deep learning, self-supervision, local features, endoscopy

Abstract: Routine endoscopes have been widely used in medical diagnosis. Three-dimensional (3D) modelling reconstruction of endoscopic images has become the development direction of future medical domain. Local feature extraction and matching is a key step for 3D modelling reconstruction. Handcrafted local features such as SIFT, SURF, ORB, are still a predominant tool for such tasks. Due to the special environment of endoscopes, there are generally weak textures and large lighting changes, which make traditional feature point extraction algorithms unable to extract feature points well. We explore the potential of the self-supervised method SuperPoint. Many existing works have shown the benefits of enhancing spatial encoding. We propose a new architecture unit, in which the SE attention mechanism module is proposed, which can explicitly model the interdependence between convolutional feature channels to improve the network's representation ability. The experimental results show that this multi-scale channel attention feature point extraction algorithm based on SuperPoint has better result and achieves higher matching quality than handcrafted local features and original algorithm in endoscopic images.

1. Introduction

Endoscopy is a medical procedure that allows doctors to inspect and observe the inside of the body without performing major surgery [1]. MIS(minimally invasive surgery) offers smaller incisions, faster recovery, and less noticeable scars [2]. However most endoscopes are capable of providing only 2D(two-dimensional) images/videos and rely on the physician's experience to localize the lesion’s location within the 3D space. There are significant risks during this process [3]. Human endoscopic 3D reconstruction technology based on computer vision has been proposed to solve this question. The method is to process the abdominal cavity image and establish a three-dimensional model of the abdominal cavity image to obtain surgical field of view and surgical spatial depth. This process could prevent surgical errors and improve patient safety [4]. The important aspect in image-based 3D reconstruction is to extract and match feature points from abdominal images [5].

Our work is to study abdominal images and find suitable algorithms for extracting abdominal feature points. We conducted experiments on traditional feature point extraction and deep learning

feature point extraction methods. Comparative experimental results, found handmade local feature calculations perform poorly on abdominal images. Images in the human abdominal cavity are easily affected by shadows, overexposure, and difficult to extract smooth feature points in the special abdominal environment [6]. It is important to find a feature point extraction algorithm and feature point descriptor that can be well used in the abdominal endoscopic environment [7]. SuperPoint [8] is one of the seminal works in this topic and one of the deep learning methods that has achieved good performance in feature point extraction. Such learning-based localized features are used in the field of endoscopic images. We have studied the algorithm process and proposed a feature point extraction algorithm based on attention mechanism called SE_block [9]. The mechanism models the interdependence between channels of convolutional features, the quality of representations generated by the network can be improved. Compared with traditional and original networks, the improved network achieves better results in feature point extraction and matching.

2. Introduction to relevant algorithms

Based on the types of features extracted, current digital feature extraction methods can be divided into two categories[10]. One is based on completely handcrafted design features, it is mainly based on mathematical calculations. The other is to learn and extract features through deep learning.

The first one mainly extracts information by calculating and abstracting images through mathematical formulas, some representative algorithm include SIFT(Scale-invariant Feature Transform) [11], SURF(Speeded Up Robust Features) [12] and ORB(Oriented FAST and Rotated BRIEF) [13], etc.

For feature point extract, The SIFT algorithm first constructs a Gaussian pyramid on the image by blurring and subsampling the images sequentially. Then it enhances the features by subtracting of one blurred version of an original image from another, then search extreme points in DoG(Difference of Gaussian) and second-order Taylor polynomial is performed on these extreme points. Keypoints are identified as local extrema in scale-space within the DoG pyramid. The descriptor determine the most significant direction of change by analyzing the gradient size and direction in the local neighborhood around feature points. This ensures that the descriptor remains consistent even if the feature point is rotated in the image. SURF upgrades over SIFT, the primary change of the SURF lies in its ability to rapidly compute operators through the utilization of box filters. Box filters can be swiftly computed using integral images. Integral images allow for the rapid calculation of sums of pixel values over rectangular regions, regardless of filter size, reduces computational complexity in feature descriptor computation. ORB mixes the FAST(Features from Accelerated Segment Test)[14] keypoint detection and the BRIEF(Binary Robust Independent Elementary Features) [15] descriptor. FAST is an intensity-based corner detection algorithm used to identify keypoints. It compares the intensity of pixels in a circular region around a candidate pixel to determine whether it's a corner. BRIEF is a feature descriptor that generates binary feature descriptors for these keypoints. It generates compact binary descriptors for keypoints by comparing the intensity of pixel pairs within a local patch around each keypoint. Rotational invariance achieved through vectors between geometric centers and grayscale centroids. Although feature extraction based on fully handcrafted design has been developed for many years, handcrafted feature point and descriptor extraction algorithms extract information from images through mathematical formulas, which inherently lack the robustness and generalization compared to large-scale data-driven deep learning approaches. Especially in the process of extracting and matching feature points for specific scenes, handcrafted methods have poor feature extraction capabilities and may not be adaptable to specific tasks.

Therefore, using deep learning methods for feature point extraction or feature descriptor computation has become a new direction of development. Unlike traditional feature point extraction

methods, utilizing deep learning-based feature point extraction allows for adaptive adjustments in feature point extraction for various specific scenes and imaging conditions. Due to the better robustness and generalization of deep learning, the feature points or descriptors obtained in this way also exhibit better performance for specific scenarios than manually designed methods. Deep learning feature point extraction algorithms have flourished in recent years, Tian proposed L2-Net [16], which utilizes convolutional neural networks to learn features of image patch descriptors in Euclidean space. The network takes down sampled image blocks of size 32 ×32 as input and outputs a 128-bit floating-point descriptor. The network describes the distances between feature points using the L2 norm and employs a loss function to constrain the matched image patches. Although L2-Net can access a large number of training samples, this can easily lead to overfitting and result in poor results. The SuperPoint [8] proposed by DeTone et al. It is a typical example of end-to-end learning, which inputs an image and outputs feature points and feature point descriptors. A self supervised approach is proposed to train the network to extract feature points and descriptors. The network structure consists of a shared encoder and two branches. Shared encoder is for extracting image features. The extracted features are input into two branches, which respectively output feature points and feature point descriptors.

Through big data-driven deep learning, deeper image features can be extracted than manually designed algorithms, and it has stronger robustness. This article focuses on studying feature point extraction networks based on attention mechanism to address the current challenges faced by endoscopic 3D reconstruction. With the dataset of abdominal endoscopy, the network structure is optimized to enable deep networks to extract good feature points from abdominal images in special endoscopic environments, promoting the development of 3D reconstruction of human endoscopes.

3. Self-Supervised Feature Point Detection Algorithm with Attention Network

3.1. SuperPoint introduction

SuperPoint is an end-to-end feature point and descriptor extraction network. The input is a complete image, and a shared encoder is used to extract deep features from the input image and reduce the dimensionality of the output image. The shared encoder network is a VGG-like network structure, including convolutional layers, max pooling layers, and nonlinear activation layers. The image's width and height are reduced to one-eighth of their original size through three max-pooling layers. After passing through the shared encoder, the input changes from $I \in \mathbb{R}^{H\times W\times 1}$ (H and W are the height and width) to $\mathcal{B} \in \mathbb{R}^{H\times W\times 1}$ with smaller spatial dimension and greater channel depth. Afterward, the output is passed through both the keypoint decoder and the descriptor decoder. The output of the keypoint decoder is the probability that each pixel in the image is a keypoint. The output of the descriptor decoder is a descriptor of feature points, and the process is to use a network similar to UCN(Universal Correspondence Network) to obtain a semi-dense descriptor, which reduces the training memory cost of the algorithm while reducing the running time. In the end, The feature point descriptor transform into a tensor $\mathcal{D} \in \mathbb{R}^{H/8\times W/8\times 256}$.

In the training of the network, MagicPoint is first used as the base detector and train it through a synthetic geometric dataset. This dataset mainly consists of simple 2D geometric shapes, including triangles, quadrilaterals, ellipses, straight lines, etc. This base detector evidently possesses certain detection capabilities for prominent keypoints. However, it overlooks many potential interest point locations. Its repeatability across multiple viewpoints is inadequate, and lacks generalization to effectively extract keypoints in real image. Therefore, self-supervised training and homographic adaptation are proposed. First, the images are subjected to homographic transformations. Then, MagicPoint is used to generate a set of pseudo ground truth for each image, followed by learning

using conventional supervised learning mechanisms. The most important thing in this process is the homographic Adaptation, which performs homographic transformations. For a given image, its keypoints should not vary with changes in viewpoint. Homographic transformations are applied to the images to obtain pseudo variations in viewpoint. The coordinates of keypoints can then be easily mapped to corresponding points in the transformed images using the homography transformation formula. Therefore, utilizing supervised learning in this context makes it easy for the network to acquire the ability to extract keypoints from different perspectives. The combination of MagicPoint and homographic adaptation detector improves the performance of the detector and generates pseudo real points of interest. The resulting detection results are more reproducible. The detector obtained at this time is named SuperPoint.

Recent studies have shown that integrating learning mechanisms into networks can help capture spatial correlations between features and enhance network representation. One approach is to integrate multi-scale processes into network modules to achieve better performance and enable better modeling of spatial dependencies. A new construction module called SE (Squeeze Excitation) [9] block is introduced, which aims to improve the quality of representations generated by the network by explicitly modeling the interdependence between channels of convolutional characteristics. This network allows for feature recalibration, enabling it to learn and utilize global information while selectively emphasizing informative features and suppressing irrelevant ones. Its structure involves using a branch to learn and assess the correlation information between different channels, which is then applied to the original feature map to achieve input recalibration. This branch aids in learning representations that are more suitable for the neural network. To enable the network to obtain channel correlations through global information, global pooling is used in the architecture to capture global information. Subsequently, two fully connected layers are connected to process the input, completing the input recalibration process and allowing the network to learn better representations. SE attention mechanism as depicted in Figure 1.

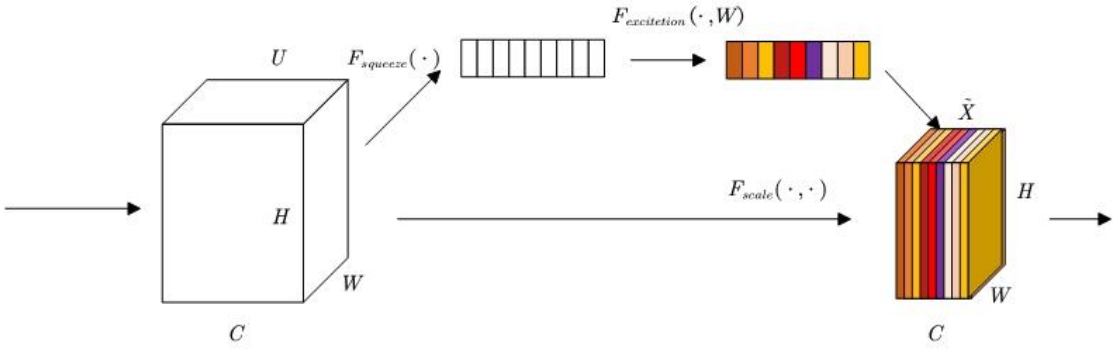

Figure 1: SE attention mechanism.

The input is squeezed by $F_{squeeze}(\cdot)$ and excited by $F_{excitetion}(\cdot, W)$. This process could module captures the correlation between convolutional feature.

3.2. SE attention mechanism based SuperPoint

The shared encoder network undergoes a Squeeze operation, compressing the spatial dimensions of the features, converting each dimensional feature channel into a scalar. This scalar possesses a global receptive field to some extent, and its output dimension matches the input feature channel count. It characterizes the global distribution of responses on feature channels, and enables layers close to the input to also obtain a global receptive field. Next is the Excitation operation, which is a mechanism similar to a gate in recurrent neural networks. We need to generate weights for each feature channel through parameters, where the parameters are learned to model the correlation

between feature channels. Finally, there is a Reweight operation that considers the weight of the output of the Excitation as the importance of each feature channel after feature selection. Then, by multiplying each channel, it is weighted onto the previous features, completing the recalibration of the original features in the channel dimension.

Based on the characteristics of the attention mechanism, we modify the encoder network, enabling the network to extract multi-scale features from the input image. The image is convolved and outputs a tensor with 64 channels, then, the attention network is used to extract features from this tensor. It generatea weights for each feature channel by analyzing the different channels. The parameters are learned to model the correlation between feature channels. The network goes through squeeze to explore channel dependencies for each channel. But each of the learned filters operates with a local receptive field and consequently each unit of the transformation output is unable to exploit contextual information outside of this region. To make use of the information aggregated in the squeeze operation, the network goes through excitation. The excitation operator maps the input specific descriptor to a set of channel weights. The relationships are not confined to the local receptive field the convolutional filters are responsive to. After that, the features extracted by the attention network are sent to the next layer of the network. In the sharing encoder, it passes through convolutional layers, max pooling layers, and nonlinear activation layers. The size of the image reduces to $H_c = H/8$ and $W_c = W/8$ through three max pooling layers. The process as depicted in Figure 2.

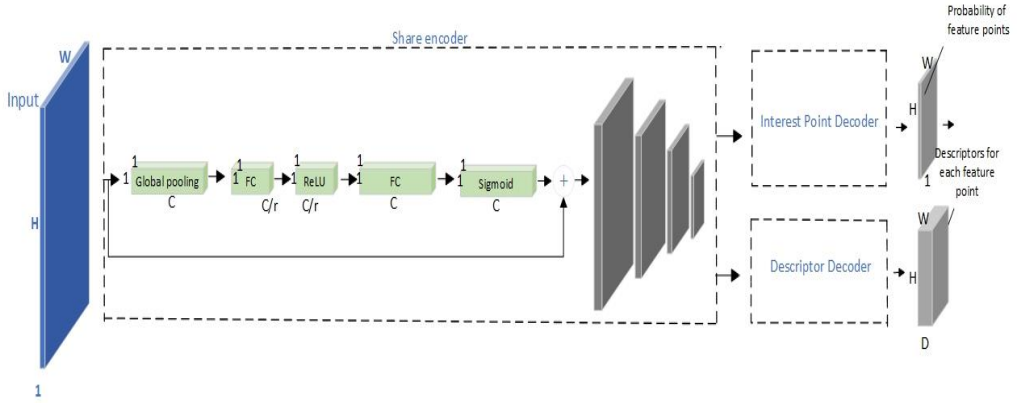

Figure 2: SE attention mechanical framework.

3.3. Feature point matching

Image matching is an important evaluation index for feature point extraction networks. To compare the superiority of algorithms, we use the same feature point matching method as traditional feature point extraction methods and deep learning methods, which is bidirectional brute force matching. There is an error in the bidirectional brute force matching, RANSAC(Random Sample Consensus) conducted geometric estimation to remove outliers. The exploration was focused on extracting and matching feature points between consecutive frames of the video, the structure and deformation of the abdominal cavity were not considered.

4. Experiment

Dataset: There are two batches of data used, the EndoSLAM [17] dataset and the HyperKvasir dataset[18].

The EndoSLAM dataset is a comprehensive dataset in the field of simultaneous endoscopic localization and map construction (SLAM) research, covering 3D point cloud data, capsule and standard endoscopic video recordings, and synthetic data for alignment or detection of six porcine

organs. A total of 39406 frames are for training and validation.

HyperKvasir is the largest image and video dataset of the gastrointestinal tract available today. The data is collected during real gastro- and colonoscopy examinations at Bærum Hospital in Norway and partly labeled by experienced gastrointestinal endoscopists. The dataset contains 110,079 images and 374 videos, and represents anatomical landmarks as well as pathological and normal findings. The total number of images and video frames together is around 1 million. The HyperKvasir dataset can play a valuable role in developing better algorithms and computer-assisted examination systems for gastro and colonoscopy.

Training set: The goal of this article is to perform good feature point extraction on the abdominal cavity, which is an important preliminary work for SLAM (Simultaneous Localization and Mapping) 3D reconstruction. The SLAM can be described as: a robot starts moving from an unknown position in an unknown environment, and during the movement, it locates itself based on the position and map. At the same time, an incremental map is constructed on the basis of its own localization to achieve autonomous localization and navigation of the robot. This means that there should be some differences between the training set and the test set. In this article, EndoSLAM is used as the training set, and the video frames in EndoSLAM are extracted as the training set. A total of 39406 frames of images are extracted from the videos, of which 31524 frames are randomly selected as the training dataset.

Validation set: Use randomly selected 7882 frames from EndoSLAM as the validation set, with a training validation set ratio of 8:2.

Test Set: Videos from the HyperKvasir dataset were selected as the test set. The selection criteria were videos with minimal surgical instrument interference and annotations that had little impact on image information. A total of 125 videos were selected from the dataset, each consisting of 5 seconds of video footage randomly. This resulted in the extraction of 15,740 frames in total. (Note: Some videos have different frame rates, either 25 frames per second or 30 frames per second.)

Evaluation indicators: Accuracy and repeatability are used to evaluate for two different images.

Figure 3 and Figure 4 show feature extracted and matches from neighboring frames.

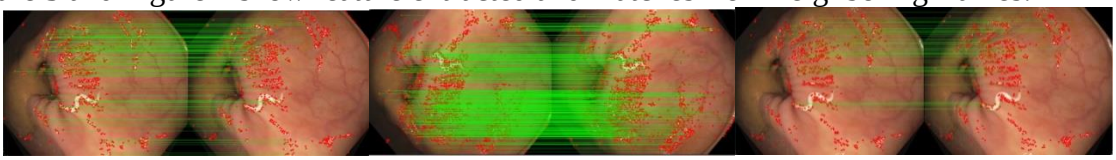

Figure 3: Feature extraction (red circle) and matching (green line) on endoscopy samples, SIFT, SURF, ORB.

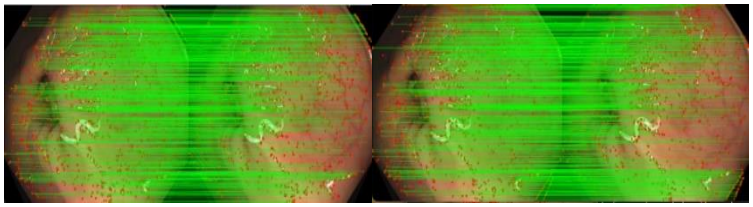

Figure 4: Feature extraction (red circle) and matching (green line) on endoscopy samples, SuperPoint, OURS.

As shown in the Figure 3 and Figure 4, there is a significant gap between hand-crafted local features and deep learning-based ones. In hand-crafted local features, the surf extracts the most feature points and successfully matches the most feature points as well. But they are all worse than the results of deep learning-based ones. These pictures show the advantages of deep learning feature point extraction. The results of the dataset are shown in Table 1.

Table 1: Result table.

Method	Average number of feature points per pair of images	Average number of correctly matched feature point pairs	Repeatability[19](%)	Accuracy (%)
SIFT	647.14	91.24	39.18	68.04
SURF	1364.73	238.75	41.20	58.39
ORB	548.31	54.52	33.92	60.25
SuperPoint	966.80	250.29	58.16	72.69
OURS	1192.04	304.97	58.70	75.62

Average number of feature points per pair of images: It is the average value of all feature points in two images matched with feature points.

Average number of correctly matched feature point pairs: It is the average logarithm of all feature point matches in two images that match feature points.

$$repeatability = \frac{N_c}{min(N_b,N_a)} \tag{1}$$

$$accuracy = \frac{N_c}{N_m} \tag{2}$$

Repeatability/Accuracy: Calculation method such as formula 1 and 2. N_b and N_a respectively represent the number of feature points extracted from adjacent frame images. N_c represent the number of correct match feature points. N_m represent the number of force match feature points. In each match, we can get a Repeatability and Accuracy, average all reproducibility and accuracy.

As shown in the table, we find that our method has advantages in terms of feature point extraction quantity and correct matching. Repeatability and accuracy are higher than the original method.

5. Conclusion

This work investigates the performance of feature point extraction in endoscopic environments and compare the effect of handcrafted local features with deep learning ones. Although handcrafted local features are widely used, we find that deep learning has great advantages over handcrafted local features in feature point extraction and matching in the abdominal environment. Improved feature point extraction and matching by adding attention modules in deep networks. They mainly learn and use global information, while selectively emphasizing information features and suppressing useless features in the graph. This improves its feature point extraction ability. The number, repeatability and accuracy of feature point extraction are all higher than the original method.

References

[1] H. Dubois, J. Creutzfeldt, M. Törnqvist, and M. Bergenmar, 2020, "Patient participation in gastrointestinal endoscopy—From patients' perspectives," Health Expectations, 23, 4, 893-903.

[2] A. Darzi and Y. Munz, 2004, "The impact of minimally invasive surgical techniques," Annu. Rev. Med., 55, 223-237.

[3] F. Chadebecq, F. Vasconcelos, E. Mazomenos, and D. Stoyanov, 2020, "Computer vision in the surgical operating room," Visceral Medicine, 36, 6, 456-462.

[4] C. Fergo, J. Burcharth, H.-C. Pommergaard, N. Kildebro, and J. Rosenberg, 2017, "Three-dimensional laparoscopy vs 2-dimensional laparoscopy with high-definition technology for abdominal surgery: a systematic review," The American Journal of Surgery, 213, 1, 159-170.

[5] A. Zaman, F. Yangyu, M. Irfan, M. S. Ayub, L. Guoyun, and L. Shiya, 2022, "LifelongGlue: Keypoint matching for 3D reconstruction with continual neural networks," Expert Systems with Applications, 195, 116613.

[6] O. L. Barbed, F. Chadebecq, J. Morlana, J. M. Montiel, and A. C. Murillo, 2022, "Superpoint features in endoscopy," in MICCAI Workshop on Imaging Systems for GI Endoscopy, 45-55.

[7] X. Liu et al., 2020, "Extremely dense point correspondences using a learned feature descriptor," in Proceedings of

the IEEE/CVF conference on computer vision and pattern recognition, 4847-4856.
[8] D. DeTone, T. Malisiewicz, and A. Rabinovich, 2018, "Superpoint: Self-supervised interest point detection and description," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 224-236.
[9] J. Hu, L. Shen, and G. Sun, 2018, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 7132-7141.
[10] Q. Liu, J. Zhang, J. Liu, and Z. Yang, 2022, "Feature extraction and classification algorithm, which one is more essential? An experimental study on a specific task of vibration signal diagnosis," International Journal of Machine Learning and Cybernetics, 13, 6, 1685-1696.
[11] A. Witkin, 1984, "Scale-space filtering: A new approach to multi-scale description," in ICASSP'84. IEEE international conference on acoustics, speech, and signal processing, 9, 150-153.
[12] H. Bay, T. Tuytelaars, and L. Van Gool, 2006, "Surf: Speeded up robust features," in Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, 404-417.
[13] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, 2011, "ORB: An efficient alternative to SIFT or SURF," in 2011 International conference on computer vision, 2564-2571.
[14] D. G. Viswanathan, 2009, "Features from accelerated segment test (fast)," in Proceedings of the 10th workshop on image analysis for multimedia interactive services, London, UK, 6-8.
[15] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, 2010, "Brief: Binary robust independent elementary features," in Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece 778-792.
[16] Y. Tian, B. Fan, and F. Wu, 2017, "L2-net: Deep learning of discriminative patch descriptor in euclidean space," in Proceedings of the IEEE conference on computer vision and pattern recognition, 661-669.
[17] K. B. Ozyoruk et al., 2021, "EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos," Medical image analysis, 71, 102058.
[18] H. Borgli et al., 2020, "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," Scientific data, 7, 1, 283.
[19] K. Mikolajczyk and C. Schmid, 2005, "A performance evaluation of local descriptors," IEEE transactions on pattern analysis machine intelligence, 27, 10, 1615-1630.