Efficient Image Analysis for Low Quality Medical Images

Chen Gong

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Eric J. Seibel, Chair

Steven L. Brunton

Blake Hannaford

Jonathan T.C. Liu

Program Authorized to Offer Degree:
Mechanical Engineering

University of Washington

Abstract

Efficient Image Analysis for Low Quality Medical Images

Chen Gong

Chair of the Supervisory Committee:
Professor Eric J. Seibel
Mechanical Engineering

Medical images captured with low-cost devices have some similar low quality characteristics such as a small field of view (FOV) compared to the overall size of the imaged area, low or degraded spatial resolution, and repetitive background with sparse features. Different from images of general scenarios with many distinctive features, these low quality medical images need more targeted approaches for analysis. This study focuses on the 2D medical images captured from different low-cost imaging devices. The medical images and videos are analyzed with two types of feature learning: one is based on the unsupervised dimension reduction and another one is based on the supervised neural network. In the first type, the coarse-to-fine frameworks for different downstream image analysis tasks are proposed. As the coarse step, the dimension reduction learns the embedding of input images in a low dimensional space which allows more efficient and reliable computation. It provides an efficient initialization for different methods in the following fine step, which narrows their optimization or search domain. The developed coarse-to-fine frameworks outperform classic methods considerably in template matching, mosaicking and localization tasks over low quality medical images and videos, including synthetic data, phantom data and *in vivo* data. It is the first proposed two-stage framework leveraging dimension reduction for efficient analysis of low-quality images, which can also be generalized into different downstream tasks by modifying the fine stage and achieve the state-of-the-art (SOTA) performance.

In the second type, the deep learning based analysis is studied in the real-time eye tracking by localizing small FOV retina frames with large distortion. After backbone of the convolution neural network, the feature map of the input frame is designed as the kernel to perform convolution on the feature map of the full retina in the following convolution layers. To achieve a more robust performance under the frame distortion, a Kalman filter is used to embed the deep learning in the state transition model and the feature-point based template matching as the measurement. Combining neural network with feature point matching to reduce the noise in two process, the motion distortion is eliminated in the tracking result which outperforms relying on neural network only. A robotic platform is built to collect annotated dataset for supervised training of the network. Considering the overfitting to the noisy labels during training, the approach of learning with noisy labeling is explored. The proposed framework combines the idea of small loss selection and noise label correction, which learns network parameters and reassigns ground truth labels iteratively. The proposed method achieves the SOTA performance on both synthesized and real noise labelled dataset. It is independent to the backbone network structure and can be directly used in the training of different models by adding a Siamese network. The Siamese network is removed after training, the original model then can be used for test.

ACKNOWLEDGMENTS

My greatest thanks are to my advisor, Eric Seibel, who has been both a role model and a source of encouragement and guidance throughout my time at University of Washington. Eric provides me the opportunity to achieve my goal to contribute on the medical field with my engineering skills, which is the start of my career journey. He always remembered what I wanted to do and helped to find any possible opportunities for me to achieve my goals. I explored many medical applications under his guidance including the medical robot I am most interested in, and he was always supportive and responded quickly during my PhD time. With Eric's recommendation, I interned at Intuitive Surgical and got more deeply involved in to the surgical robotics that originally inspired me in this path. As an advisor, his support extends well beyond research. He often encouraged me to try more things and ask for what I want. This made me more confident to stand out. He is also caringly concerned my life, and it makes me feel at home when I study abroad alone. I would not have what I achieved my highest goals without Eric's help and he is my advisor for life.

I am also thankful for Steven Brunton, my co-advisor. He is very inspiring for my development of algorithms and math in different applications and very supportive in every one of our meetings. He is knowledgeable and intelligent and provided a large amount of good ideas and broadened my eyes during my study. The well-organized classes he lectured solidified my mathematics which was required for my research. I also appreciate his suggestions of how to write good technical papers and learned a lot from the writing booklet he sent me.

I thank Professor Blake Hannaford and Professor Jonathan Liu for being on my thesis committee and for advice along the way. I am grateful for Dr. Brian Schowengerdt and Dr Laura Trutoiu from Magic Leap providing funding and advice on my research. I would also

like to thank Weisi Xie, Yuanyuan Shi and Yaxuan Zhou for the conversations and supports during my PhD as good friends.

Working in Human Photonics Lab has truly been a pleasure. I enjoyed the time spent with the current and former labmates: Cathy Olivo, Matt Carson, Andy Lewis, Yaxuan Zhou, Yang Jiang, Pengcheng Chen, Len Nelson, Manuja Sharma, and Shawn Swanson to name a few. It is great to collaborate and chat with them regularly and I was taught a great deal.

All of my love and thanks go to my parents who were my first and great teachers. Thanks to my younger brother who accompanied Mom when I am abroad and being so independent and considerate. I look forward to reuniting with them soon after the pandemic.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Motivation for low quality medical image analysis

The growth of use and sophistication in computer vision technologies brings advances in the computer-assisted processing and analysis of images, but medical images are different and more challenging than the natural images captured from the everyday scenes. Focusing on the 2D optical imaging, most of the image analysis downstreaming tasks, such as image classification, detection and registration, are based on recognizing the image features from coarse-grained to fine-grained levels. The natural images or videos generally contain rich information or features which can contribute to the analysis, and the objects to be recognized often take a large percentage of the image. However, medical images of tissues typically have repeating low contrast background with sparse features as the key information (for example, sparse vessels and small lesion area) because of the human tissues' similarity. This results in a much lower signal-to-noise ratio when recognizing the key features. In these conditions, recognition and analysis of the key features in medical images is more difficult and more vulnerable to noise than natural images, which is one of the main reasons that many computer vision tasks for natural images or videos are considered to be fully solved while remain challenging in the medical field.[4] On the other hand, local changes are highly possible during medical imaging or between different times of imaging the same tissue regions. Some examples are the dynamic processes of specular reflections, bubble formation, media light scattering changes, motion artifacts, physical tissue deformations and natural growth of tissue and lesion appearance over time.

Nowadays, a large amount of low cost and portable devices are developed for medical imaging, with the expansion of the medical needs in rural, economical, nurse-operated, highly

distributed primary care facilities. However, with reducing the size and cost of the device, the image quality is generally held to be inferior to those of standard devices, in terms of the FOV, resolution, distortion and noise level. For example, in ophthalmology, portable low-cost fundus cameras become increasingly popular. The potential of smartphone retinal imaging in teleophthalmology is also growing. Currently, the smartphone based fundus imaging (SBFI) has a small FOV, limited at 30° or 45° with pupil dilation, which is 2.5 times smaller than the conventional fundus camera.[6,7] Due to the smaller FOV, examination of the retina periphery is more laborious with SBFI compounded by the fact that the eye is in constant motion. If it is compensated by a lens with a greater FOV, image quality will be reduced (reduced image sharpness and increased reflexes) and examination technique become more challenging.[6] Another example is in ultrathin flexible endoscopes which traditionally have a limited number of pixels. A new design based on a micro-optical scanner reduces the tradeoff between FOV and resolution which exists with small scope size (limited pixels). The scanning fiber endoscope (SFE) is a low-cost device with miniature probe size and large FOV compared to conventional flexible endoscopes.[8,9] This technology produces images by active scanning of laser light instead of camera imaging using pixel-array light collection from diffuse white-light illumination. However, the pixels in one SFE video frame are not scanned at the same time, which leads to image distortion based on the scanning pattern when their is relative motion between SFE and the target. The extra movement distortion further increase the analysis difficulty on images with insufficient distinctive features.[10]

Fig. 1.1 shows several examples from different medical applications (image sources and references are described in the caption), containing the specific challenges in image processing and analysis. What they have in common are the small FOV and sparse features. The low resolution and blur are more obvious in (a), (b) and (c). (d) contains bubbles as shown in the bottom left region. Besides the described challenges, the quality of medical data can be influenced by other common degradation such as specular reflection, imaging noise, motion blur and illumination changes. Driven by recent progress in machine learning and deep learning communities, analyzing medical data with neural networks (NN) has been under

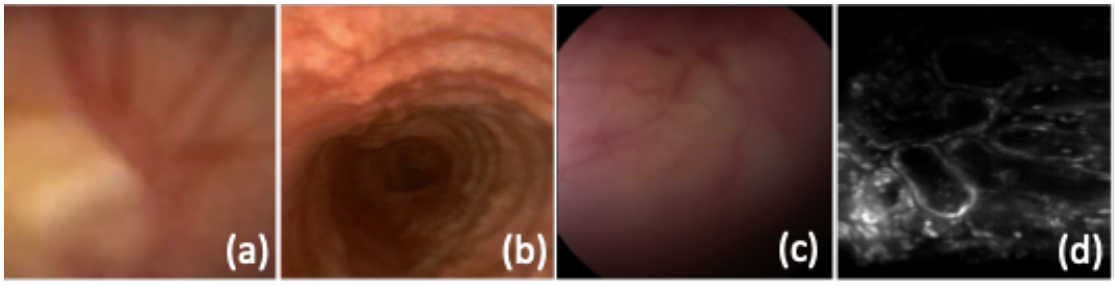


Figure 1.1: Medical image examples with challenges for analysis. (a) is the retina images captured by the smartphone fundus imaging.[1] (b) is the image of the proximal third of the trachea obtained with *Ambu aScope 4*.[2] (c) is the porcine bladder imaged with Karl Storz (Tuttlingen, Germany) HD-view flexible digital cystoscope performed in our lab. (d) is the florescence image of mouse kidney imaged by a handheld confocal microscope.[3]

active research.[11] In medical field, the quantity and quality of the data labels required for supervised training of the model can also be a limitation.[12,13] These difficulties of computer-assisted analysis of 2D medical images motivate us to develop image analysis algorithms and frameworks to provide corresponding solutions.

1.2 Organization and contribution

The principal contribution of this work is to provide a number of algorithm solutions to different down-streaming medical data analysis tasks with low data quality. The algorithms can be divided into analysis with unsupervised low dimensional features and with deep features in NN. In the first category, a generic coarse-to-fine analysis framework is developed utilizing low-dimensional models as an efficient feature extractor in the coarse step, and different fine steps can be customized according to the down-streaming tasks. In the second category, the algorithm combining deep learning models and classic image analysis are designed for the human tissue data with large distortions. A deep learning model structure and training strategy are proposed for robust learning in noisy labelled dataset. The applications in this study are based on 2D medical image analysis, whereas the developed algorithms can be modified for 3D cases in a straight-forward manner.

Chapter 2 describes the analysis with unsupervised low dimensional features in a coarse-to-fine framework. This includes the template matching of retinal images captured with smartphone fundus imaging used for tele-ophthalmology 2.1; The efficient intensity-based mosaicking framework of low quality images 2.2; Real time camera localization during robot-assisted telecystoscopy 2.3. In addition, more clinical applications in ophthalmology besides tele-ophthalmology are explored driven by the low dimensional features in 2.1.

Chapter 3 presents dealing with low quality data with supervised deep learning models. A framework combining the real-time localization neural network with the classic image registration designed for the motion distortion is proposed in 3.1. In addition, A robotic data collection platform to collect annotated data for supervised training of the model is described. In 3.2, a synergistic network learning and label correction model is provided for label noise robust network training.

Chapter 4 discusses the future work of using deep learning features in the real time camera localization for tele-cyctoscopy 4.1, and embedding the image distortion characteristics into and end-to-end deep learning model.

Chapter 2

ANALYSIS WITH UNSUPERVISED LOW DIMENSIONAL FEATURES

2.1 RetinaMatch: efficient template Matching of Retina Images

2.1.1 Introduction

Telemedicine applications are emerging at a rapid pace due to innovations in hardware and software, and changing attitudes of clinicians, providers and consumers. Teleophthalmology is an important component of telemedicine, and it is now arguably the standard of care in linking patients in remote areas to ophthalmologists. Recently, low-cost teleophthalmology has been facilitated by smartphone-based fundus imaging. In addition, the emerging virtual and mixed reality sector may enable new teleophthalmology scenarios for long-term eye imaging and monitoring. However, in the case of portable fundus photography, non-mydriatic image quality is more vulnerable to distortions, such as uneven illumination, noise, blur and low contrast.[14] In this paper, we address the challenging problem of automated retinal image matching and registration to enable future teleophthalmology applications.

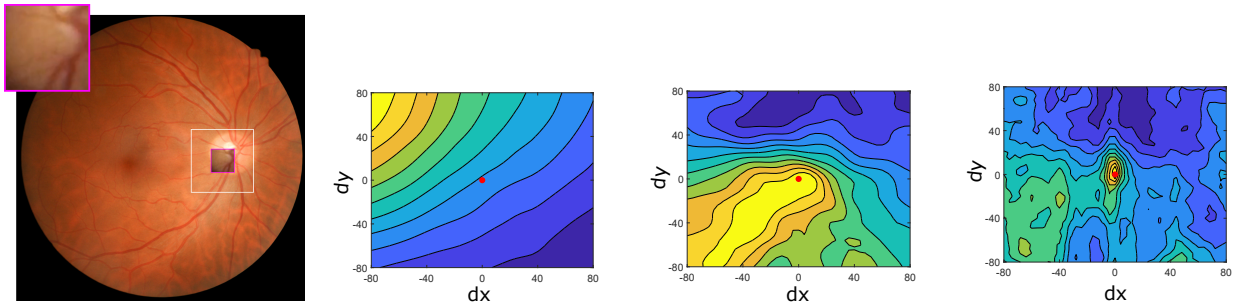
Motivation

The eye provides a unique opportunity to image internal biological tissue in vivo and many diseases can be diagnosed and monitored through ocular imaging. For example, diabetic retinopathy is a common retinal complication associated with diabetes, causing microaneurysms, exudates and hemorrhages on the retina.[15,16] Changes of retinal arteries and veins, as well as their ratios, can be indicators of hypertension.[17] The timely detection of these pathological changes via regular retinal screening and analysis is particularly important for early diagnosis and prevention.

High-quality fundus images of the retina are traditionally acquired in a laboratory setting with expensive and cumbersome equipments. Acquiring high-quality fundus images poses a significant challenge for patients in rural and other underserved areas who must overcome significant hurdles to receive regular checkups in the clinic. Visiting an ophthalmologist is often inconvenient for patients in the city as well. In contrast, emerging portable and low-cost fundus cameras allow fast, accessible imaging of the retina, albeit with a decrease in image quality. Using portable fundus cameras outside the clinic connects rural patients with their doctors.[18,19] By daily retinal monitoring and trend analysis of the data, ocular disease may no longer be considered the silent disease, as early onset is likely to be detectable and even predicted.[20]

A typical example of a portable fundus camera involves a clip-on lens adapter attached to a smart-phone system.[18] These consumer-grade optical devices have two main disadvantages: small FOV and lower image quality than lab-based fundus cameras. The FOVs of current clip-on lenses range from 5° to 20° in undilated eyes.[18,21] In this case, many small images captured in the undilated eye at different locations are necessary to obtain adequate retinal imaging. The same retinal locations need to be re-imaged and matched in order to monitor changes longitudinally over time. Accordingly, all of the captured small FOV images can be registered and compared to a stored wide FOV retinal image. This reference image is a baseline which can be stitched together by a series of small FOV images, or can leverage wider FOV images captured from a conventional ophthalmoscope. Taking the small FOV images as the templates to be matched, it is a template matching process, as shown in Fig. 1(a). The template only covers a small area on the retina, thus is unlikely to be affected much by the nonlinear deformation due to the non-planar eyeball surface. The location of the template in the full image may be represented by an affine linear transformation, i.e. including translation, rotation, shear, and scaling. This provides a mathematical framework to formulate template matching as an optimization problem.

As described above, an accurate template matching method to deal with small FOV and low quality template images is needed in teleophthalmology. Since the method will be



(a) Template matching sample (b) SSD (c) NCC (d) NMI

Figure 2.1: Alignment functions with respect to translations between the template and the white boxed area. The full FOV image in (a) is taken with the fundus camera in the clinic. The top left image in the magenta square is the template captured by a typical adapter-based fundus camera *D-eye*, having only translations along two axes. (b)(c)(d) show the alignment function between the template and regions within the white boxed area. The true alignment position is $(0, 0)$ – see red dots. Only NMI shows an obvious maximum at the alignment position. Note that the optimal value of SSD is minimum and NCC and NMI are maximums.

implemented on portable devices, the efficiency of computation is also a driving requirement.

Related Work

Much of the foundational work on template matching of retinal images is based on more general image registration methods, which have been comprehensively studied in recent years. However, general retina registration methods focus on matching image pairs that both have a large FOV with local deformations or different image modalities. Many existing retinal template matching algorithms are limited to detecting specific objects from the image, where the template always contains a certain feature, such as the optic disc, exudate and artifacts.[22–24]

Retinal image registration itself is challenging: the nonvascular surface of retina is homogeneous in healthy retinas, while exhibiting a variety of pathologies in unhealthy retinas.[25] Retina images captured by adapter-based optics provide less information and have low image quality, which further increases the difficulty of template matching. It is instructive to introduce current retina image registration methods which can be used for template matching and

their feasibility in addressing our stated problem. Retina image registration approaches can be classified into area-based and feature-based methods. Feature-based methods optimize the correspondence between extracted salient objects in retina images.[25–29] Bifurcations, fovea, and the optic disc are common features used for retinal image registration. A small FOV template has little probability of containing specific landmarks on the retina, thus the fovea and optic disc are not applicable. Vascular bifurcations are more common, while similarly, the small number of bifurcations in the template cannot form the basis of a robust registration. Besides, the extraction of the vascular network in poor quality images is difficult. General feature point based approaches are also implemented in retina registration, such as SIFT-based[30,31] and SURF-based methods.[32,33] These approaches can register the images in complex scenarios and are computationally efficient. They assume the feature point pairs can be reliably detected and matched to estimate the transformation. Although feasible in most cases, the process can fail on low-quality retina images without enough distinct features.

Area-based approaches match the intensity differences of an image pair under a similarity measure, such as SSD (sum of squared differences),[34] CC (Cross-Correlation)[35] and MI (mutual information),[36] then optimize the similarity measure by searching in the transformation space. Avoiding pixel-level feature detection, such approaches are more robust to poor quality images than feature-based approaches. However, retina images with sparse features and similar backgrounds are likely to lead the optimization into local extrema. Fig. 2.1 shows an example of the area-based method with three similarity measures. The small template image is captured by the adapter-based *D-eye* optics which is registered onto a full fundus image. Both of the images are acquired by the same modality. SSD and normalized CC (NCC) do not have an obvious peak at the alignment position (0,0), giving no clear information on the alignment quality. Normalized MI (NMI) shows a maximum at the alignment position, while it still has local extrema which can interfere with the global optimization. Besides, when the size difference between the template and full image is too large, registration with MI can be computationally prohibitive.

Contributions

In this paper, we present RetinaMatch, a new template matching method that overcomes the challenges posed by registering small FOV and low-quality retinal images onto a full image. This approach is an improvement over the area-based methods that only optimize the MI metric,[36] since it achieves more accurate and robust template matching near the alignment position, as shown in Fig. 2.1.

The unique aspect of our approach is that we combine dimension reduction methods with the MI-based registration to reduce the sensitivity to local minima, while improving the matching efficiency. An overview of our novel template matching framework is shown in Fig. 3.13. Specifically, the principal component analysis (PCA) and block PCA are used to localize the template image coarsely, then the resulting displacement parameters are used to initialize the MI metric optimization. The initial parameters provided by the coarse localization are in the convergence domain of MI metric. In this way, the transformation search space for optimization is narrowed significantly. The PCA computation is accelerated with randomized methods,[37–39] which improves the coarse localization efficiency. Both the use of PCA for coarse localization and the use of randomized methods for acceleration are unique methods of implementation. Further, we have carefully compared PCA against several other dimension reduction techniques, and we find that PCA offers the best tradeoff in simplicity, accuracy, and efficiency. Another contribution is that this paper proposes an efficient image mosaicking algorithm based on the image dimension reduction. It accelerates the matching of overlapped images among unordered data, especially in image mosaicking with area-based registration methods.

The proposed method is validated on the STARE retinal dataset[40] with synthetic deformations, and *in vivo* data captured by a low-cost (<US$400) adapter-based optical device *D-eye*. The performance of different dimension reduction techniques are also compared on the STARE dataset. RetinaMatch can find the correct mapping even when the image is of poor quality with non-distinct features, whereas other methods fail due to unstable feature

detection and local extrema.

2.1.2 Preliminaries

PCA for Location Estimation

Dimension reduction methods allow the construction of low-dimensional summaries, while eliminating redundancies and noise in the data. To estimate the template location in the 2d space, the full image dimension is redundant, thus we apply dimension reduction methods for the template coarse localization. In this section we describe the dimension reduction methods we use in this paper.

Generally, we can categorize dimension reduction techniques as either linear or nonlinear. The most prominent linear technique is principal component analysis (PCA), which dates back to the work of[41] and.[42] PCA is selected as the dimension reduction method in RetinaMatch since it is simple and versatile. Specifically, PCA forms a set of new variables as a weighted linear combination of the input variables. Consider a matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_d]$ of dimension $n \times d$, where n denotes the number of observations and d is the number of variables. Further, we assume that the matrix $\mathbf{X}$ is column-wise mean centered. The idea of PCA is to form a set of uncorrelated new variables (so called principal components) as a linear combination of the input variables:

$$\mathbf{z}_i = \mathbf{X}\mathbf{w}_i, \tag{2.1}$$

where $\mathbf{z}_i$ is the ith principal component (PC) and $\mathbf{w}_i$ is the weight vector. The first PC explains most of the variation in the data, the subsequent PCs then account for the remaining variation in descending order. Thereby, PCA imposes the constraint that the weight vectors are orthogonal. This problem can be expressed compactly as the following minimization:

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{X} - \mathbf{Z}\mathbf{W}\|_F^2 \\ &\text{subject to} \quad \mathbf{W}^\top\mathbf{W} = \mathbf{I}, \end{aligned} \tag{2.2}$$

where $\|.\|_F$ is the Frobenius norm. The weight matrix $\mathbf{W}$ that maps the input data to a subspace turns out to be the right singular vectors of the input matrix $\mathbf{X}$. Often a low-rank approximation is desirable, i.e., we compute only the k dominant PCs using a truncated weight matrix $\mathbf{W}_k = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_k]$.

PCA is generally computed by the singular value decomposition (SVD). Many algorithms have been developed to streamline the computation of the SVD and PCA for high-dimensional data that exhibits low-dimensional patterns.[43] In particular, tremendous strides have been made to accelerate the SVD and related computations using randomized methods for linear algebra.[37–39] Since we have demonstrated high performance with less than 20 principal components, the randomized SVD is used to compute the principal components, improving the efficiency in this retinal mapping application for mobile platforms. The randomized algorithm proceeds by forming a sketch $\mathbf{Y}$ of the input matrix

$$\mathbf{Y} = \mathbf{X}\mathbf{\Omega}, \tag{2.3}$$

where $\mathbf{\Omega}$ is a $d \times l$ random test matrix, say with independent and identically distributed random standard normal entries. Thus, the l columns of $\mathbf{Y}$ are formed as a randomly weighted linear combination of the columns of the input matrix, providing a basis for the column space of $\mathbf{X}$. Note, that l is chosen to be slightly larger than the desired number of principal components. Next, we form an orthonormal basis $\mathbf{Q}$ using the QR decomposition $\mathbf{Y} = \mathbf{QR}$. Now, we use this basis matrix to project the input data matrix to low-dimensional space

$$\mathbf{B} = \mathbf{Q}^\top \mathbf{X}. \tag{2.4}$$

This smaller matrix $\mathbf{B}$ of dimension $l \times d$ can then be used to efficiently compute the low-rank SVD and subsequently the dominant principal components. Given the SVD of $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, we obtain the approximate principal components as

$$\mathbf{Z} = \mathbf{QU}\mathbf{\Sigma} = \mathbf{XV}. \tag{2.5}$$

Here, $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors and the diagonal elements of $\mathbf{\Sigma}$ are the corresponding singular values. The approximation accuracy can be controlled via additional oversampling and power iterations, for details see.[44]

It is important to note that PCA is sensitive to outliers, occlusions, and corruption in the data. In ophthalmological imaging applications, there are several potential sources of corruption and outliers when imaging the full image, including blur, uncorrected astigmatism, inhomogeneous illumination, glare from crystalline lens opacity, internal reflections (e.g., from the vitreoretinal interface and lens), transient floaters in the vitreous, and shot noise in the camera. Further, there is always a trade-off between illumination and image quality, and there is strong motivation to introduce as little light as necessary for the patient comfort and health. The robust principal component analysis (RPCA)[45,46] was introduced specifically to address this issue, decomposing a data matrix into the sum of a matrix containing low-rank coherent structure and a sparse matrix of outliers and corrupt entries. In general, RPCA is more expensive than PCA, requiring an iterative optimization to decompose the original matrix into sparse and low-rank components. Each step of the iteration is as expensive as regular PCA, and typically on the order of tens of iterations are required; however, PCA may be viewed as an offline step in our procedure, so that this additional computational cost is manageable. RPCA has been applied with success in retinal imaging applications to improve image quality.[47,48] In the examples presented in this work, the data appears to have few enough outliers so that RPCA is not necessary, although it is important to keep RPCA as an option for data with outliers and corruption.

Mutual Information

In this section, we describe the maximization of MI for multimodal image registration. We define images $\mathbf{S}$ and $\widehat{\mathbf{S}}$ as the template and target images, respectively. A transform u is defined to map pixel locations $x \in \mathbf{S}$ to pixel locations in $\widehat{\mathbf{S}}$.

The main idea of the registration is to find a deformation $\widehat{u}$ at each pixel location x that maximizes the MI between the deformed template image $\mathbf{S}(u(x))$ and the target image $\widehat{\mathbf{S}}(\mathbf{x})$.

Accordingly,

$$u_{opt} = \arg\min_{u} MI(\mathbf{S}(u(x)), \widehat{\mathbf{S}}(x)), \tag{2.6}$$

where

$$MI(\mathbf{S}(u(x)), \widehat{\mathbf{S}}(x)) = \sum_{i_1 \in S} \sum_{i_2 \in \widehat{S}} p(i_1, i_2) log(\frac{p(i_1, i_2)}{p(i_1)p(i_2)}). \tag{2.7}$$

Here, i_1 and i_2 are the image intensity values in $\mathbf{S}(u(x))$ and $\widehat{\mathbf{S}}(\mathbf{x})$, respectively, and $p(i_1)$ and $p(i_2)$ are their marginal probability distributions while $p(i_1, i_2)$ is their joint probability distribution.

2.1.3 Proposed Approach

In this section, we describe RetinaMatch, which combines dimensionality reduction and mutual information based image registration. From Fig. 2.1 we can see MI performs better than other similarity metrics even on the same modality images, thus we focus on the MI criterion. Given a large FOV full image and a small FOV template image, our method can be used to localize the template on the full image accurately and efficiently. The full image can be a wide-field fundus image or a mosaicked one from D-eye images. The underlying concept is to use PCA and block PCA first for coarse localization, which can be a warm start to following accurate registration. In accurate registration, the MI metric is optimized to find the optimal transformation. Since the optimization domain has been narrowed to a small range near the optimal position with coarse localization, the accurate registration can achieve high accuracy and efficiency. Fig. 2.2 provides an overview of the general approach to RetinaMatch.

2.1.4 Coarse Localization with Dimension Reduction

We define the full image and the template as $\mathbf{F}$ and $\mathbf{S}$ respectively. The full image $\mathbf{F}$ is split into target images $\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_N$:

$$\mathbf{I}_i = \phi(b_i, \mathbf{F}). \tag{2.8}$$

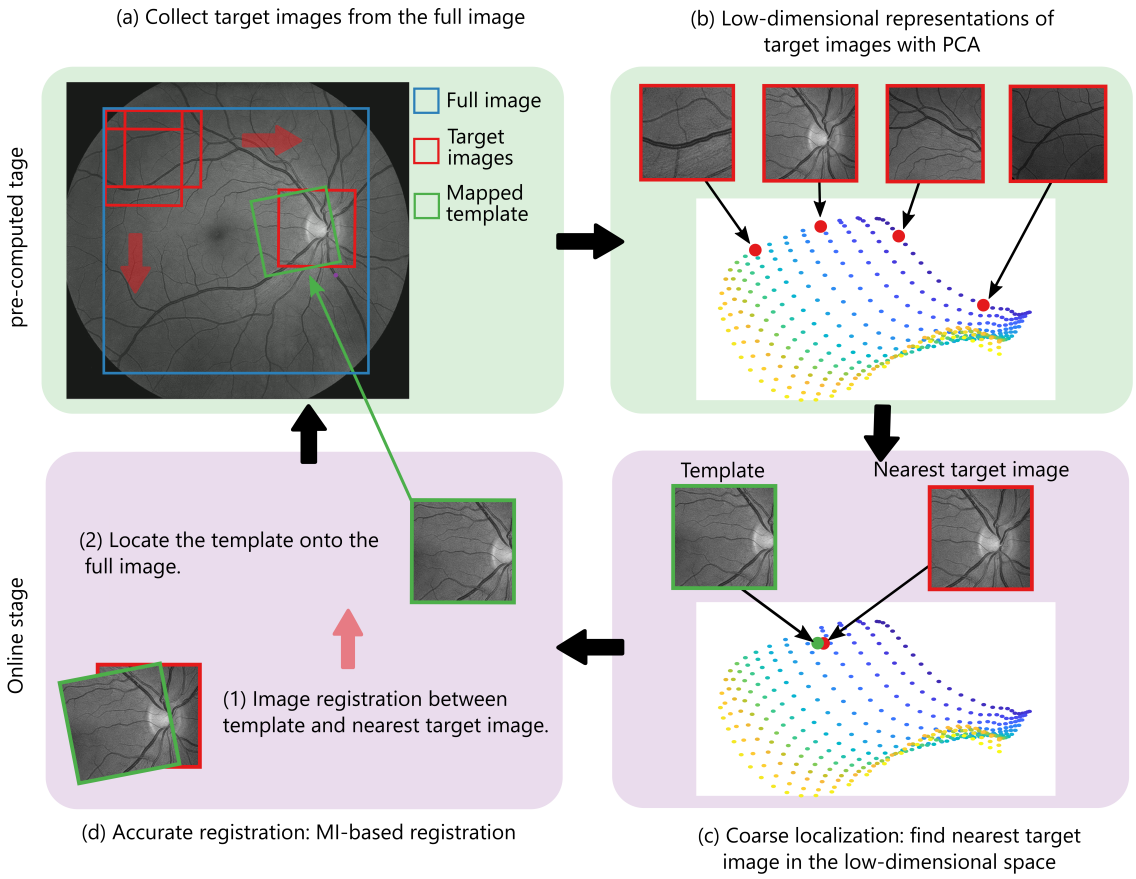


Figure 2.2: Schematic of the proposed retinal template matching method shown in four panels from (a) to (d) . In panel (a) the wide-FOV full image is sampled with many overlapping target images. (b) Each target image is mapped into the low-dimensional space according to its positional relationship. (c) An example template is also mapped into this space and its nearest target image is found. (d) The nearest target image is registered with MI. The panels (a) and (b) in green can be pre-processed offline when the full image is obtained, while panels (c) and (d) are considered as the online stage. The schematic describes the method without using the improvement of block PCA. Please see Sect. III for more details of block PCA.

The function ϕ crops $\mathbf{I}_i$ from $\mathbf{F}$ at b_i and $b_i = [x_i, y_i, h, w]$, where (x_i, y_i) denotes the center position and (h, w) denotes the width and height of the source image. There is a certain displacement f of neighboring target images in the x and y axes. As shown in Fig. 3.13(a), each target image has a large overlap with its neighbors. The overlap forms the redundancy of the data which can indicate the location distribution between each image and its neighbors. Applying dimension reduction techniques on such data we can obtain the low-dimensional distribution map of all target images.

Target images are resized to vectors and form the matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$. We obtain the low-dimensional distribution representation of the target image distribution by implementing PCA on $\mathbf{X}$:

$$\mathbf{Z} = \mathbf{X}\mathbf{W}, \tag{2.9}$$

where $\mathbf{Z} = [z_1, z_2, z_3, ..., z_N]^T \in \mathbb{R}^{n \times l}$, $\mathbf{W} \in \mathbb{R}^{d \times l}$ and $l \ll d$. The image space Ω_1 is mapped to a low-dimensional space Ω_2 with the mapping $\mathbf{W}$. $\mathbf{W}$ and $\mathbf{Z}$ are saved in the dictionary $\mathcal{D}$.

Given a template $\mathbf{S}$, the coarse position can be estimated by recognizing its nearest target image. The nearest target image in Ω_1 should also be the nearest representation of $\mathbf{S}$ in Ω_2. Accordingly, we obtain the low-dimension feature z_s of the template in Ω_2:

$$z_s = \tilde{s}\mathbf{W}, \tag{2.10}$$

where $\tilde{s} \in \mathbb{R}^d$ is the reshaped vector of template $\mathbf{S}$. Let $\Delta(z_s, z)$ be the Euclidean distance between z_s and a feature z in $\mathbf{Z}$. z^* is the nearest target feature of the source image $\mathbf{S}$ in Ω_2:

$$z^* = \arg\min_z \Delta(z_s, z). \tag{2.11}$$

The corresponding target image location is used as the coarse location of $\mathbf{S}$. Ideally, the difference between the coarse location and the ground truth in x and y axes should be less than $f/2$ pixels.

In the first experiment in Sect. IV, PCA outperforms other non-linear dimension reduction methods, while the error is larger than $f/2$. The main reason is that the image degradation creates spurious features that contribute to the final classification. To reduce the influence of local spurious features, we implement block PCA to further improve the accuracy of the coarse localization. By computing the PCA of different local patches in the template, the effect of local features, which leading to the template can not be located correctly, is reduced.

Here we introduce the detailed process of the block PCA. Obtaining the nearest target image, we crop a larger image at the same position from the full image as the new target image $\mathbf{I}$. In this way, the template can have more overlap with the new target image when there is a large offset between two images. We segment $\mathbf{I}$ and the template $\mathbf{S}$ into small patches with the function $\tilde{\phi}$, where the patch size is smaller than the source image with the axial displacement of neighboring patches f'. Similarly, all image patches from $\mathbf{I}$ are mapped into the low-dimension space Ω_3 with $\mathbf{W}'$. Let $\mathbf{Z}'$ denote the low-dimensional representation of the target image distribution. Each template patch is then mapped to the space with $\mathbf{W}'$[1]. The nearest target patch for each template patch is determined with the Euclidean distance as described before. The coordinates of each target patch represent the location of the mapped patch. We use the same weight for each region of the template for localization, thus the average of all template patches location can be taken as the template's location. Let b_m be the mean of the coordinates of selected nearest target patches, which then represents the center of the template on $\mathbf{I}$. Accordingly, the template location on the full image can be estimated and the region is cropped as the image $\widehat{\mathbf{S}}$. The accurate registration is then applied to the template $\mathbf{S}$ and image $\widehat{\mathbf{S}}$. In this way, the coarse localization provides an estimate of a good initial point for the accurate registration.

In the implementation of the proposed coarse localization, the full image is assumed to exist so the dictionary $\mathcal{D}$ can be built in advance. This is the pre-computed part as shown

[1]Fig. S1 in the supplementary material gives an example of the image patch mapping.

in Fig. 3.13 (a-b). The process after the template being acquired is called the online stage, involving the block PCA for coarse localization followed by the accurate registration. The online stage of the coarse localization is shown in Algorithm 1.

Algorithm 1: Coarse localization: online stage

1 Map template $\mathbf{S}$ into space Ω_2: $z_s = \tilde{s}\mathbf{W}$.
2 Determine closest target image $\mathbf{I}$ with corresponding $z^* : z^* = \arg\min_z \Delta(z_s, z)$. $z^* \in \mathbf{Z}$.
3 Segment $\mathbf{S}$ into $[S_p^1, S_p^2, ..., S_p^n]$: $S_p^i = \tilde{\phi}(b_i, \mathbf{S})$; Segment $\mathbf{I}$ into $[I_p^1, I_p^2, ..., I_p^n]$: $I_p^i = \tilde{\phi}(b_i, \mathbf{I})$.
4 Map target patches I_p^i into space Ω_3: $\mathbf{Z}' = \mathbf{I_p}\mathbf{W}'$, where $\mathbf{I_p}$ is formed with vectorized I_p^i.
5 For each template patch S_p^i:
6 (i)Map S_p^i into space Ω_3: $\tilde{z_s}^i = S_p^i\mathbf{W}'$.
7 (ii)Determine its closest target patch $I_p^{Idx(i)}$ with index $Idx(i)$.
8 $b_m = \frac{1}{n}\sum_{i=1}^{n} b_{Idx(i)}$, where $b_{Idx(i)}$ is the coordinate of selected target patch $I_p^{Idx(i)}$.
9 **return** localization region $\widehat{S} = \phi(b_m, \mathbf{F})$).

Accurate Registration

In this section, images $\mathbf{S}$ and $\widehat{\mathbf{S}}$ are accurately registered with maximization of mutual information. The location of $\widehat{\mathbf{S}}$ on the full image $\mathbf{F}$ becomes the estimated displacement of the template $\mathbf{S}$. As the small FOV of template images, the relationship between the template and the full image can be modeled by linear transformations. In our work, the transform u for alignment is given as an affine transformation:

$$u = \begin{bmatrix} a_{11} & a_{12} & t_x \\ a_{21} & a_{22} & t_y \\ 1 & 1 & 0 \end{bmatrix}. \tag{2.12}$$

From the MI equation 2.16, we can see the MI function has a discrete formulation which

is not differentiable. Several solutions therefore are proposed to smooth the MI function to compute the MI derivatives and keep its accuracy. We use the method described in,[49] where the joint probability distribution between the images $\mathbf{S}$ and $\widehat{\mathbf{S}}$ is estimated with a Parzen window.

The optimizer used for the MI maximization is based on Newton's method. The MI function is a quasi-concave function (Fig. 2.1(d)), and the parabolic hypothesis of the Newton's method is only valid near the convergence. When the initial transformation is on the convex part of the cost function, it will cause the optimization to diverge. In the example of Fig. 2.1(d), the normalized MI measure has local extrema interference. The proposed coarse localization provides a good initialization of the displacement for subsequent optimization of the MI alignment function. In the figure, the estimated alignment position is $(11, 9)$. The estimation is close to the optimal value and falls in the convex domain of the MI metric, which provides more efficient optimization and avoid local extrema.

After registration between images $\mathbf{S}$ and $\widehat{\mathbf{S}}$, the template $\mathbf{S}$ can be matched on the full image $\mathbf{F}$ based on the position b_m of the selected region $\widehat{\mathbf{S}}$. Fig. 2.3 shows a schematic of the coarse localization process and intermediate results.

Image Stitching

As pointed out in Sect. I, the full retina image can be stitched into a panorama by using many small templates. Users must capture a series of images in naturally unconstrained eye positions to explore different regions of the retina. It is problematic to determine adjacent images before the registration when we apply area-based registration approaches, since they do not have effective descriptors for matching.

Related to the dimension reduction in the proposed template matching method, here we present Algorithm 2 to learn the positional relationship of images to be stitched. In this way, the adjacent images can be recognized and registered efficiently. For a series of small images $\mathbf{X_i}$, we form the matrix $\mathbf{X}$, as with the matrix $\mathbf{T}$. PCA is applied to $\mathbf{X}$ and returns the low-dimensional features for each image in Ω_2. The distance between features in Ω_2

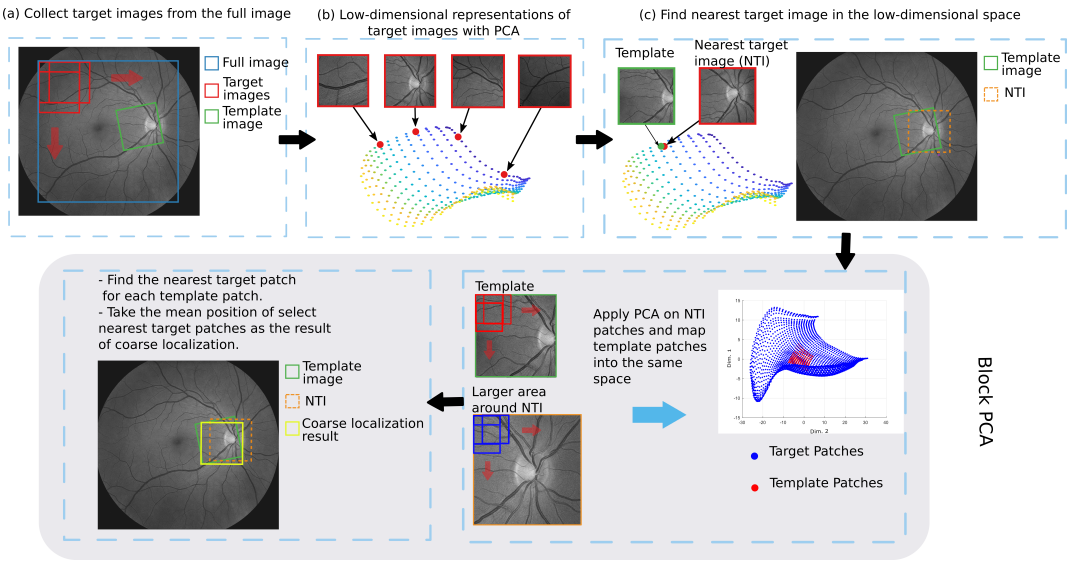


Figure 2.3: Schematic of the coarse localization process: PCA and block PCA.

indicates the distance between images. The nearest neighbor $\mathbf{X_j}$ of image $\mathbf{X_i}$ is the one with largest overlap; the image pair is then registered with MI-based approach. To improve the algorithm robustness, the 3-nearest neighbors for each image are first selected to compute MI with, and we keep the one with the largest metric value.

Algorithm 2: Image stitching

1 Map images into space Ω_2:$\mathbf{Z} = \mathbf{XW}$.
2 For each image $\mathbf{X_i}$:
3 (i).Find the nearest 3 neighbors $\mathbf{X_j}$ minimizing the feature distance $\Delta(\mathbf{Z_i}, \mathbf{Z_j})$.
4 (ii).Compute the Mutual Information between each $\mathbf{X_j}$ and $\mathbf{X_i}$ and take the adjacent image with highest MI.
5 Panorama $\mathbf{R}$ Mosaicking: Align all the adjacent images with mutual information based registration method.
6 Panorama blending.
7 **return** panorama $\mathbf{R}$.

2.1.5 Experiments

We present the performance of our template matching method on three experiments using retina images. For comparison, we use the global MI algorithm described in Mattes et al.[50] and ASIFT (modified SIFT for affine deformation).[51] In the first experiment, each template is extracted from the full fundus image in the STARE dataset and matched back to it. The intermediate results of the coarse localization are also evaluated. In the second experiment, the template captured by the adapter-based optics is matched to the full fundus image captured by the clinical fundus camera. In the third experiment, a panorama is mosaicked from small templates first, and subsequently individual templates are matched to the panorama.

Fundus Images from STARE Dataset

In this experiment, we validate our method on simulated fundus images. We use images from the STARE dataset,[40] which consists of 400 raw fundus images of healthy and diseased retinas. Matching image pairs are simulated from this dataset. Each image pair includes a full fundus image selected randomly from the dataset and an affine transformation is applied to map it from a square into a parallelogram. The area within the mapped square is then cropped and warped (with the inverse affine transformation) to obtain the square template. The FOV of the template images is around 12° with a size of 200×200 pixels. The template dimension is 10% of the full image.

The ground truth is available in this experiment, thus root-mean-square (RMS) errors between corrected displacements and ground truth positions are used as a metric to measure the RetinaMatch accuracy. To evaluate the coarse localization, we take the center point distance between the template and the chosen target region.

Validation of the Coarse Localization

First the coarse localization with and without the block PCA refinement are tested. In the implementation, target images are generated with a displacement of $f = 10$ pixels and $f' = 5$

for the block PCA. We use the top 20 and 10 PCA features in the first PCA step and the block PCA respectively. The parameters are fixed in remaining experiments. Additionally, we test the coarse localization with two other non-linear dimension reduction methods: kernel PCA[52] and Isomap.[53] We compare the non-linear dimension reduction methods to see if the non-planar retina surface and the affine transformation affect the performance of the PCA-based linear method. In the kernel PCA, we compared Gaussian kernel and polynomial kernels with different degrees. The Gaussian kernel has better performance and is thus selected for kernel PCA. There may be other better kernels that better separate the data. However, finding this embedding space is labor-intensive and may need to be re-tuned for new image types, whereas PCA is more generic. The experiment is performed over 100 matching image pairs created from the STARE dataset. The pixel-level errors (coarse localization error as described), success rates, and average runtimes of these methods are shown in Table I. The criterion of successful matches in the coarse localization is a pixel-level error of less than 40 pixels. It is verified that the PCA based coarse localization is more efficient, accurate and interpretable. Block PCA further improves the accuracy while the time spent is higher than PCA-only method. To further improve the online efficiency, the target patches mapping can be precomputed for each target images. The average time spent in this case will decrease to 0.0975s.

Additional experiments were carried out to test the proposed coarse localization under different image degradations. Five degradation types in five levels are considered as follows (images are in double format $\in [0, 1]$): affine transform with the rotation/shear parameter of {5°/0.1,10°/0.2,15°/0.2,15°/0.3,20°/0.3}; additive Gaussian noise with standard deviation varied from 10% to 50% of the pixel value; image blurring with Gaussian kernels with standard deviation of {0.5,1,1.5,2,2.5}; intensity changes of {4%,8%,12%,16%,20%} of graylevels in the image, which is the nonlinear brightness change; add artificial pathological features of 1-5 levels with increasing amount and size, such as the spot of exudate (bright spots), hemorrhage (dark spots) and vessel width changes (enlarge/shrink vessel regions). Fig. 2.4 provides examples of the highest level degradation in each sequence. For each sequence and

Table 2.1: Comparison of coarse localization with different dimension reduction methods.

	Mean errors	Success rate	Runtime
Kernel PCA	57	83%	0.7035s
Isomap	27	94%	2.3634s
PCA	14	100%	0.0065s
Block PCA	8	100%	0.6143s

Table 2.2: Success rates of coarse localization per degradation level.

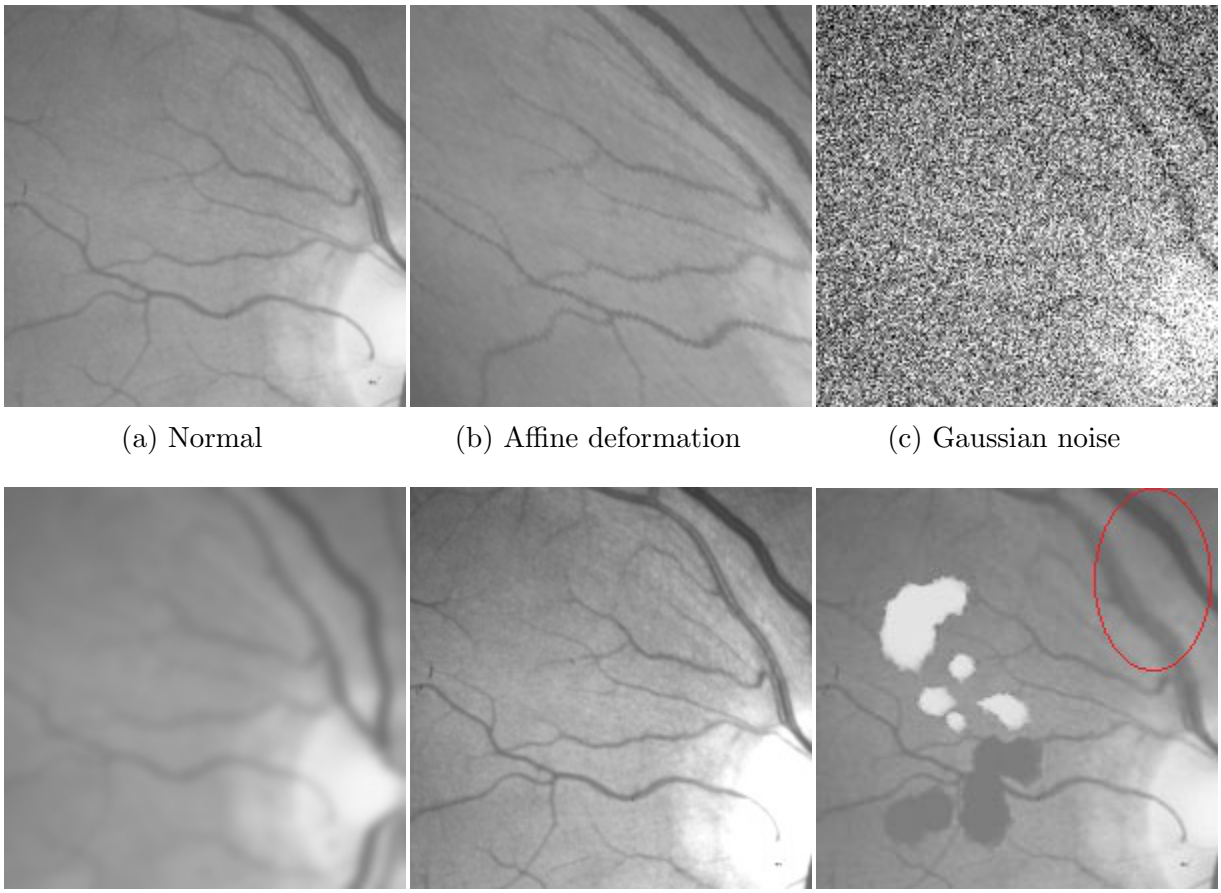
Distortion level	1	2	3	4	5
Affine deformation	100%	99%	95%	81%	75%
Noise	100%	100%	100%	100%	100%
Blur	100%	100%	100%	100%	100%
Brightness change	100%	100%	100%	100%	97%
Artificial features	100%	100%	100%	100%	98%

degradation level, we create 100 matching image pairs as described above. All degradations are applied to the template in each pair. Fig. S2[2] shows template examples of the highest degradation in each sequence. The coarse localization achieves high success rates across the dataset in different degradations, with the exception of the highest level of linear deformation sequence. However, the limitation to smaller eye rotation angle is physiologically based. The human eye has a limited range of torsional rotation with respect to the visual axis.[54] Checking a large set of data, we find the real affine deformation in the adapter-based optics imaging is less than level three (rotation/shear: 15°/0.2).

Validation of the Template Matching

We examine RetinaMatch's final performance under the same sequences and degradations described earlier, but with two additional template matching approaches: feature-based ASIFT and global MI registration. The success rate of different methods are presented in Fig. 2.5, where the successful matches are that the RMS error is less than 8 pixels. The accuracy of ASIFT decreases significantly at higher degradation levels of noise, blur, and

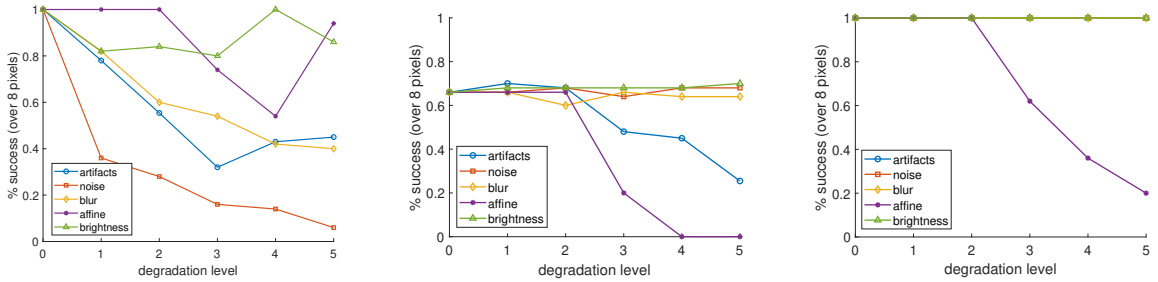
[2]Fig. S2 is in the supplementary material.



(a) Normal (b) Affine deformation (c) Gaussian noise

(d) Blur (e) Brightness (f) Artificial features

Figure 2.4: Examples of highest level degradations in each sequence. Please note that (b) is the template generated with affine deformation thus the image content is not the same. In the artificial features (f), the bright and dark spots simulate the exudate and hemorrhage, respectively. The width of vessels in the circled area is enlarged.



(a) ASIFT (b) Global MI (c) RetinaMatch

Figure 2.5: Performance of template matching methods under different image degradations. In each, the x-axis stands for the increasing levels of image degradation, ranging from 0 (no degradation) to 5 (highest). The y-axis stands for the percentage of successful matches with RMS error less than 8 pixels. All degradations have 100% success rate in RetinaMatch except three high-level affine deformations.

artifacts, due to feature-point instability. The global MI registration method cannot always converge to the correct affine transformation parameters using such small templates, thus it has a low success rate even without degradations. The performance further declines in high-level degradations of artifacts and affine. RetinaMatch has a success rate of 100% in most sequences and degradation levels, except the high-level affine deformation. As described above, the real-world affine deformation would be less than level 3. The improvement of RetinaMatch efficiency over global MI depends on the size difference between two matched images. In this experiment, the average runtime of RetinaMatch is around 50% less than that of global MI, since it narrows the search domain of the MI optimization significantly. The computation of feature points in ASIFT is expensive and it takes four times longer than RetinaMatch. ASIFT is selected for comparison because it can generate more robust feature points than SIFT.

In vivo D-eye Data and Full Fundus Image

D-eye is a typical adapter-based optical system which can convert the digital camera on smartphones to a fundus camera (https://www.d-eyecare.com/). Fig. 2.6 show several examples of D-eye images. The relatively small FOV of D-eye can be useful to monitor

the retinal health over time with comparison to a wide FOV baseline image taken at the ophthalmology clinic. With our algorithmic approach, the captured D-eye images can be matched onto the full image for automatic comparison. The latest data with retina changes can also replace the original area on the full image, maintaining a record of longitudinal changes. In this way, it offers the opportunity for a quick overall retina analysis outside the clinic, with automatic diagnostic approaches such as described in.[55,56]

This experiment is a case study with a series of D-eye data captured from one person with a healthy retina. We converted the iphone 6 to a fundus camera with D-eye, then collected the data in a dim room to provide a larger pupil and proper image contrast. The eyeball was free to rotate which allowed us to obtain images covering different regions of the retina. The collected *D-eye* images have an average FOV of 4° and a resolution of about 50 pixels/degree. The full fundus image is taken with a Kowa Nonmyd alpha-D III retinal camera, as shown in Fig. 1(a). It has a 45° FOV with a resolution of 75 pixels/degree. The D-eye images are around 0.7% of the full image. Captured with different devices, the brightness and contrast varies greatly between the image pair to be matched.

We first validate our method by matching 100 *D-eye* template images onto the full image. The ASIFT and global MI methods are also implemented. Additionally, we add pathological artificial features on the 100 D-eye templates to test the algorithm robustness to retina pathological changes. The accuracy of the template matching is evaluated using target registration error (TRE).[57] For each template, four corresponding landmarks are selected by an trained observer, two trained observers then selected the corresponding landmarks from the full image independently. To obtain TRE for each image pair, we compute the root mean square of the distance between the transformed landmark points and the landmarks selected by two trained observers. The TRE results of RetinaMatch (coarse localization and final results), ASIFT and global MI are shown in Table 2.3. Table 2.3 lists the success rate, the mean and standard deviation of TRE of successful matches and inter-observer variability. The success rate is the percentage of successful matching pairs with TRE less than 6 pixels. RetinaMatch can reach an accuracy of less than 4-pixel TRE with the observer variability,

Table 2.3: Target registration error (TRE) of template matching methods in experiment 2.

	ASIFT		MI		RetinaMatch		Observer
	Mean±SD	Success Rate	Mean±SD	Success Rate	Mean±SD	Success Rate	Variability
Without artifacts	NA	0	NA	0	3.88±1.72	94%	2.39±1.93
With artifacts	NA	0	NA	0	3.97±1.64	94%	2.39±1.93

while the ASIFT and global MI cannot match the *D-eye* image successfully.

In vivo D-eye Data and Mosaicked Full Image

In this experiment we match the D-eye templates onto the full image mosaicked with D-eye images. Using the stitched panoramic image allows the use of this device at home without going to the clinic for the full fundus image as the baseline. Inhabitants of remote areas without local eye clinics having professional fundus camera facilities can benefit greatly from this technique.

Full Image Mosaicking

The full image in this experiment is mosaicked with 20 D-eye images using the proposed image stitching method. Based on no training for the D-eye user and other limitations of the procedure, we collected images covering the region around the optic disc. In the implementation, we used the first 20 dimensions of the features when computing the image distances in the low-dimensional space. Fig. 2.6 illustrates the distribution of the first three dimensions of the features. From the two examples in the figure, we can see the nearest three neighbors of the selected sample in the low dimensional space also have a large overlap in the image space. In the image patch registration, the MI-based registration method is applied. The last row of Fig. 2.7 shows the mosaicking result with the proposed method. The MI of the top three candidate neighbors are validated to be effective to selecting the correct neighbor. The stitched full image has a 10° FOV with the same resolution as the D-eye templates. The image blending is not our focus in this paper and the mosaicked image is not perfectly seamless.

Template Matching

Table 2.4: Target registration error (TRE) of template matching methods in experiment 3. The success rate of ASIFT is 0 because the template contain an inadequate number of features. The success rate of MI is also low because the optimization gets stuck in local minima. In contrast, RetinaMatch has a consistently high success rate.

	ASIFT		MI		RetinaMatch		Observer
	Mean±SD	Success Rate	Mean±SD	Success Rate	Mean±SD	Success Rate	Variability
Without artifacts	NA	0	2.77±1.54	15%	3.06±1.65	96%	2.80±1.02
With artifacts	NA	0	3.34±1.52	8%	3.24±1.75	94%	2.80±1.02

Similar to experiment 2, we validate our method by matching 100 D-eye templates with and without pathological artificial features onto the mosaicked image. The images used for the mosaicking are not contained in the 100 template test set. The TRE results are shown in Table 2.4. The TRE of successful matches is less than 8 pixels. RetinaMatch can match 96% of image pairs without artifacts and 94% of image pairs with artifacts. The TRE results of success matches were not much different from the observer variability. On the other hand, ASIFT cannot find the alignment position since the detected feature points are not sufficient for matching. The MI approach has a low rate of success as well, which has a high probability to cause mis-detection of emerging changes. Fig. 2.7 shows examples of RetinaMatch matching results with and without artifacts in the second and third experiments.

2.1.6 Conclusion and Discussion

We present a new template matching method, RetinaMatch, which can be used in remote retina health monitoring with affordable imaging devices. A PCA-based coarse localization method is proposed to provide a good initialization for the MI-based registration in the template matching. In this way, RetinaMatch can obtain an accurate affine transformation between the image pair with poor quality and large size difference. As demonstrated in the simulation experiment, RetinaMatch does not handle templates with large affine deformations, with the success rate decreasing at level 4 and 5 in Fig. 2.5. Importantly, the template

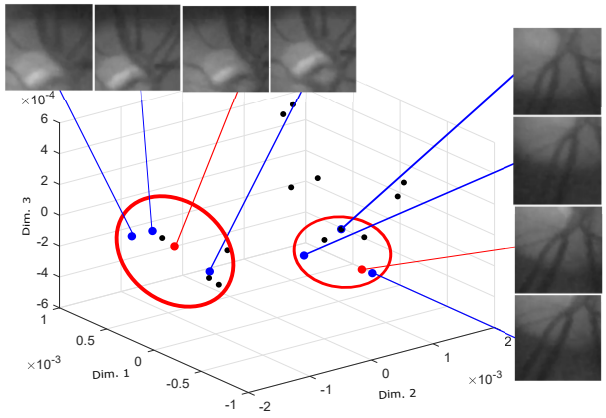


Figure 2.6: Features of images to be stitched in the top three dimensional space. Each small black dot indicates one mapped image. The colored dots in red circles show two selected samples (red) with their nearest three neighbors (blue). Note the distance is measured in the top 20 dimensional space.

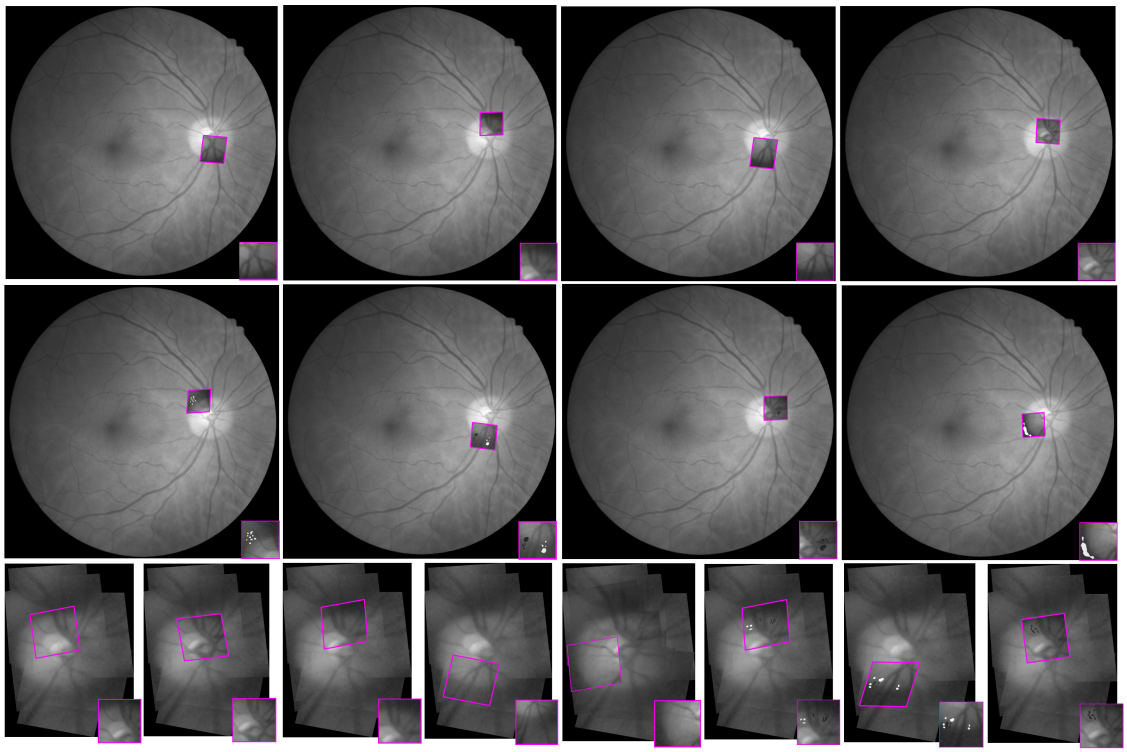


Figure 2.7: Examples of RetinaMatch results with and without artifacts. The first two rows are results of experiment 2 and the third row is experiment 3. The mapped template on the full image covers the original area and is boxed with magenta lines.

image captured by adapter-based optics with general operation will not have a linear deformation exceeding the RetinaMatch limit, therefore we can ignore the poor performance over the third-level affine degradation. To our knowledge, this is the first report addressing template matching in retina images whose template contains unconstrained small retinal areas rather than a specific object. Further algorithm testing is needed on the smartphone or other low cost fundus imaging platforms as all current testing has been limited to a PC workstation.

To validate RetinaMatch, experiments using both human datasets with simulation and *in vivo* retina images from a case study were performed. Experiments with simulated datasets allowed evaluation of the accuracy and robustness of RetinaMatch to different levels and sequences of degradations. The *in vivo* case study ensures that our method can be applied using a consumer product. It was observed that RetinaMatch provided superior performance under different image conditions over standard ASIFT and MI algorithms. The parameters, such as the offset $f = 10, f' = 5$ and the dimension $d = 20$, are independent of various datasets in the implementation of RetinaMatch, which makes it easier to translate our method to other similar imaging device besides *D-eye*.

The evaluation of the RetinaMatch accuracy is difficult in experiment 2 and 3 without ground truth. Since the goal of template matching is providing accurate alignment for downstream analysis, we use TRE as the metric. Compared with entropy based measures or the similarity measure itself, TRE measures the result intuitively in pixels and is independent of different regularization methods.

The remote monitoring of retina health with template matching is the first medical application to be proposed with RetinaMatch. Tele-ophthalmology is a promising application since many diseases are manifested at an early stage that are detectable with optical imaging of the retina. Because early stage retinal diseases do not present with symptoms, routine screening is important for early detection, which requires both high sensitivity and even higher specificity. The adapter-based optics and the digital cameras from smart phones provide an efficient and economic approach to capture retina images regularly at home. The

images of the current state can be mapped with RetinaMatch and then compared with the previous state. With regular screening, the process of lesion formation and therapeutic treatment can be monitored over time. In the experiment, *D-eye* is chosen just as one low-cost example among many others with small FOV on undilated pupils.[18] Similar fundus imaging techniques can also be implemented in emerging commercial VR, AR, and mixed reality headsets that will be widespread in the future.

There are different kinds of retina lesions that can be screened with portable fundus cameras. In the medical example of monitoring hypertension, the larger arteries constrict and the venous vessels enlarge in diameter.[17] For example, the larger blood vessel cross-sectional diameter is about 20 pixels in the case study, and a change with hypertension will be in the range of 10-60%, so we are looking for over 2-pixel changes from baseline over time. In Tables 2.3 and 2.4, the TRE is shown to be extremely low and most errors are below 2 pixels (1.8 arcmins) excluding observer variability. With advanced trend analytics,[58] we can expect template matching errors to be well below a threshold of clinical significance. For more precise vessel width measurement, RetinaMatch can be combined with vessel segmentation, as described in our previous publication.[59] The vessels of interest can be located on the current templates and the corresponding vessel width is then obtained by segmentation around the mapped location. Note the vessel segmentation here is applied on very small retina patches (20×20 pixels), which is more robust and accurate than segmentation of wide FOV retina images. The segmentation error in[59] is less than 1 pixel, which has been presented using *D-eye* images. Xu et al.[60] proposed the vessel width segmentation and measurement on retina imaging acquired from the low quality fundus camera as well. They also report similar 1-pixel accuracy. However, the imaging device they used produced higher quality retinal images, having five times larger FOV than the *D-eye*. The biomarkers of abusive head trauma (AHT) is another example. The most common retinal manifestation of AHT is multiple retinal hemorrhages in multiple layers of the retina.[61] Matching the captured images onto the full retina image, The hemorrhagic spots can be easily segmented after the subtraction of the current retina regions and previous status. The AHT then can

be recognized automatically when such spots are detected with portable fundus cameras.

RetinaMatch may be used in other medical image applications for template matching. For example, in the case of endoscopic guidance of therapy by a surgical robot,[62] the current limited-sized FOV can be matched onto the panorama for endoscope localization. Thus, this image template matching technique can be used to create a more reliable closed-loop control for the robot arm and surgical tool guidance.

2.1.7 Extension to down-streaming applications

Measurement of retinal vessel width in tele-ophthalmology for mobile health monitoring

Changes in arteriole and venule width, as well as their ratios, can be indicators of hypertension. The proposed template matching method enables the remote retina screening and measurement of retinal vessel width automatically with smartphones. The anticipated wide use of head-mounted mixed reality (MR) devices can also be equipped to provide unique opportunities to monitor eye health outside the clinic.

In the case of vessel width measurement, retinal vessels in small FOV (source) images, captured by current smartphone cameras or captured by future MR headsets, are compared to a large FOV baseline (reference) images obtained in clinic. First source images are registered to the reference image using the proposed template matching method, which has a high registration accuracy even for small FOV source images with poor image quality. Knowing measurement points on retinal vessels in reference image (pre-marked on the reference image by ophthalmologist), vessel widths of corresponding locations in source image is measured in the the following steps: A small region is cropped on the registered source image with the measurement point as the center point for retinal vessel detection. A Gaussian matched filter is used for vessel segmentation. The small region is binarized, and the detected blood vessel is thinned with a morphological operation to obtain centerline location and direction of the vessel. Finally, retinal vessel width perpendicular to the centerline around the measurement point is obtained. The pipeline is shown in Fig. 2.8.

In the experiment, four measurement points on arteries and veins are selected near the optic disk of two same retina reference images with previous experiments: one fundus image and one mosaic images from D-eye. The source images are captured with D-eye and there are 20 source images for each measurement point. Fig. 2.9 shows the results of the vessel width measurement over two reference images. In the top left figure, the red and blue dots shows the selected measurement points of arteries and veins respectively on the conventional fundus image. Top right figure shows the corresponding boxplot of the relative error of vessel width measurement with our pipeline. Most of the relative errors are within -10% - 0%. All of the 20 source images for each measurement point are registered on the reference image successfully. The measurement error is caused by vessel width measurement method, and the absolute error is around 1 pixel. Bottom left figure shows the measurement points on the mosaic reference image. The right boxplot only containing the results of three measurement point because over 50% of the source images cannot be registered successfully in the featureless region around point 2. For the rest 3 vessel points, the mean relative errors are within -15% - -5%.

These experiments test the feasibility in tele-ophthalmology using smartphones to assess retinal health. With the proposed framework, the onset of vascular disease may be detected with such a mobile device outside of the clinic. Based on the D-eye image dataset, there are still some small retinal regions can have matching errors. In this case, the proposed method can be used to monitor the vessel's width in "easy regions". Simple devices with higher image quality are expected in the future, and next generation MR and VR headsets' capable of retinal imaging could enable longitudinal monitoring of users' retina as well. Furthermore, with the proposed retinal template matching method, not only the vessel width, but any changes of retina can be monitored over time.

Optic disc tracking in SLO/OCT videos

Poor fixation or erroneous eye/head movements cause misaligned optical coherence tomography (OCT) scans and their assessment of retinal nerve fiber layer (RNFL) thickness. The

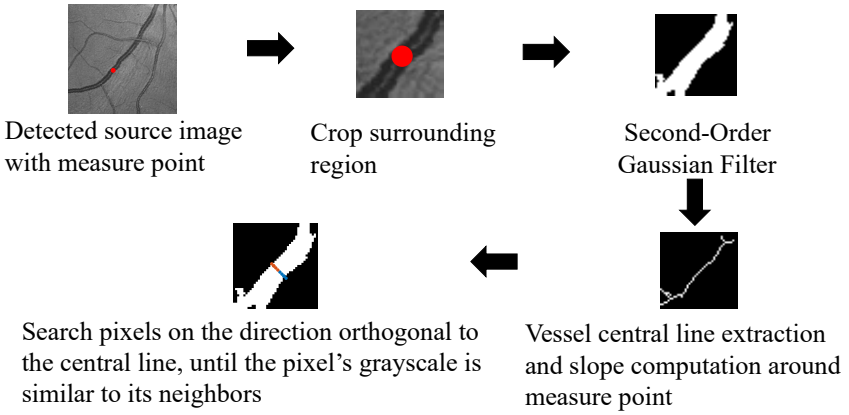


Figure 2.8: Pipeline of retinal vessel width measurement in tele-ophthalmology.

alignment artifact can make centering circular scans on the optic disc impossible. In the case of poor fixation or nystagmus, the line scans near the optic disc is typically acquired in order to get a qualitative assessment of the RNFL. Therefore, we need an accurate tracking of the optic disc in SLO/OCT videos of people with poor fixation. The dimension reduction based coarse localization can also be leveraged in this application for real time optic disc tracking.

In proposed approach, the optic disc region is manually selected from the first frame of the scanning video. For each of the remaining frame, we take the selected object as a small template and the whole frame as the image upon to map the template. The template matching is then performed in the following steps: 1. The current frame is separated into target images which have the same FOV with the template of optic disc. 2. PCA is applied on the target images set to map the target image into a low dimensional space. 3. Map the template into the same space and find its nearest target image. 4. Use SURF feature-point-based registration method to accurately register the two small patches for real-time purpose. 5. According to the position of the selected target image, map the template onto the current frame. The coarse localization improves the performance of SURF method significantly compared to the direct registration between the object and the whole frame.

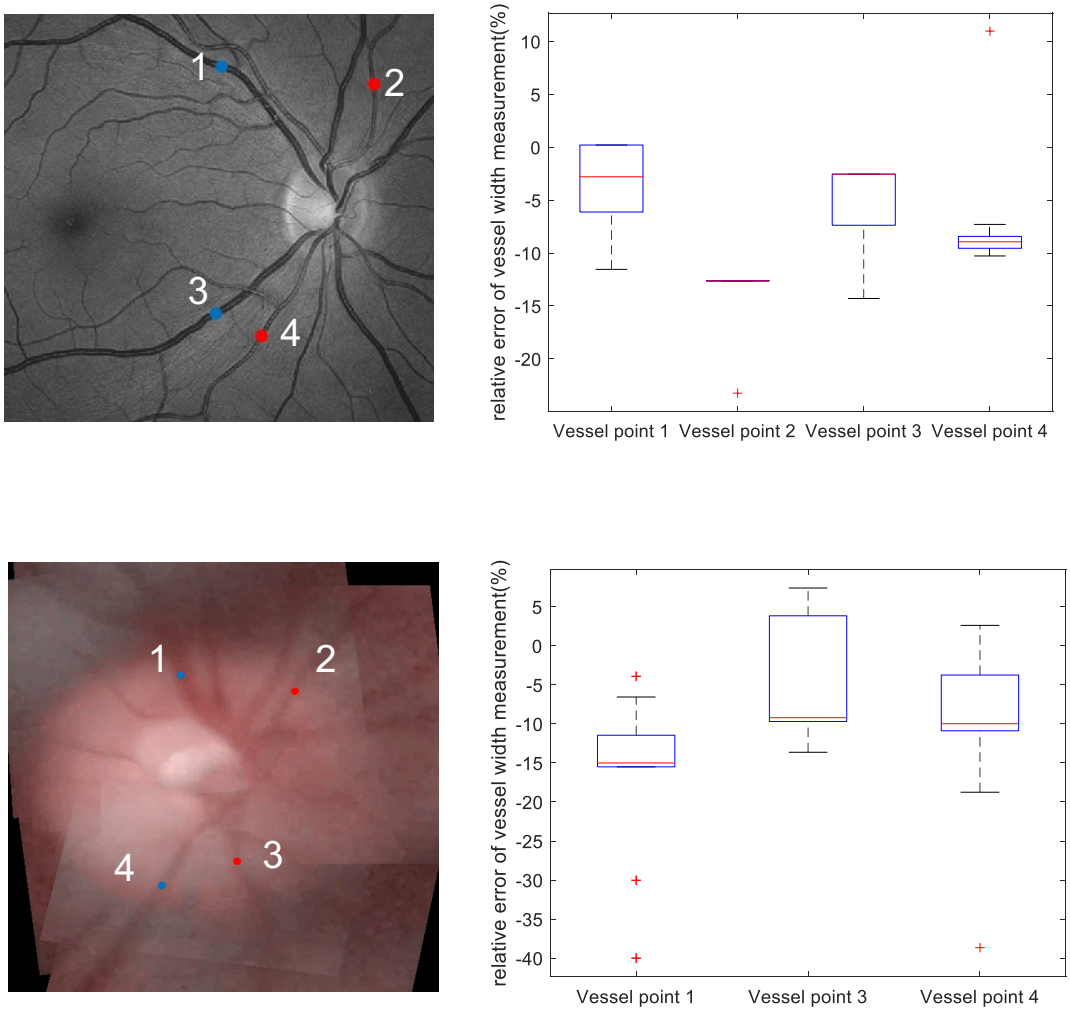


Figure 2.9: Boxplot of the vessel width measurement relative errors on two reference images.

Table 2.5: Optic disc tracking comparison between the proposed method and SURF only.

	Success rate	Mean error (Among successful cases)
SURF-only	15%	0.1182 (5.0px) ± 0.022mm
Ours	92%	0.0759 (3.2px) ± 0.017mm

Fig. 2.10 shows several examples of the optic disc tracking results from two randomly select scanning videos. The tracking is successful even with very low illumination on the frame. The last frame is a failure example because of the eye blinking. Fig. 2.11 presents two video examples of comparison between the groundtruth and tracked optic disc positions. The x and y axes means the x and y directions in mm. The tracked positions coincide with the ground truth well, and the errors contain the ground truth measurement error. The statistical comparison between the errors of our method and using SURF only is shown in Table. 2.5, where our method has a success rate of 92% while SURF-only is 15%.

Based on the result, it is validated that our method can track the optic disc accurately in a SLO/OCT video and indicate the deformation of the object. Since the detection of the optic disc in each frame is independent from previous frames, the failure cases in several low contrast frames caused by eye rotation or blinking does not influence the tracking quality of following frames. This method overperforms the general feature point based method significantly according to the success rate. The optic disc template has a very small FOV and few valid feature points. If search the optic disc globally, there is a lot of interference from other detected points in wrong regions like around the specular reflection. Our method reduced such interference significantly by obtaining a region near the optic disc location in the coarse localization. Using this pipeline, we are able to acquire the line scans near the optic disc for RNFL especially from people with poor fixation.

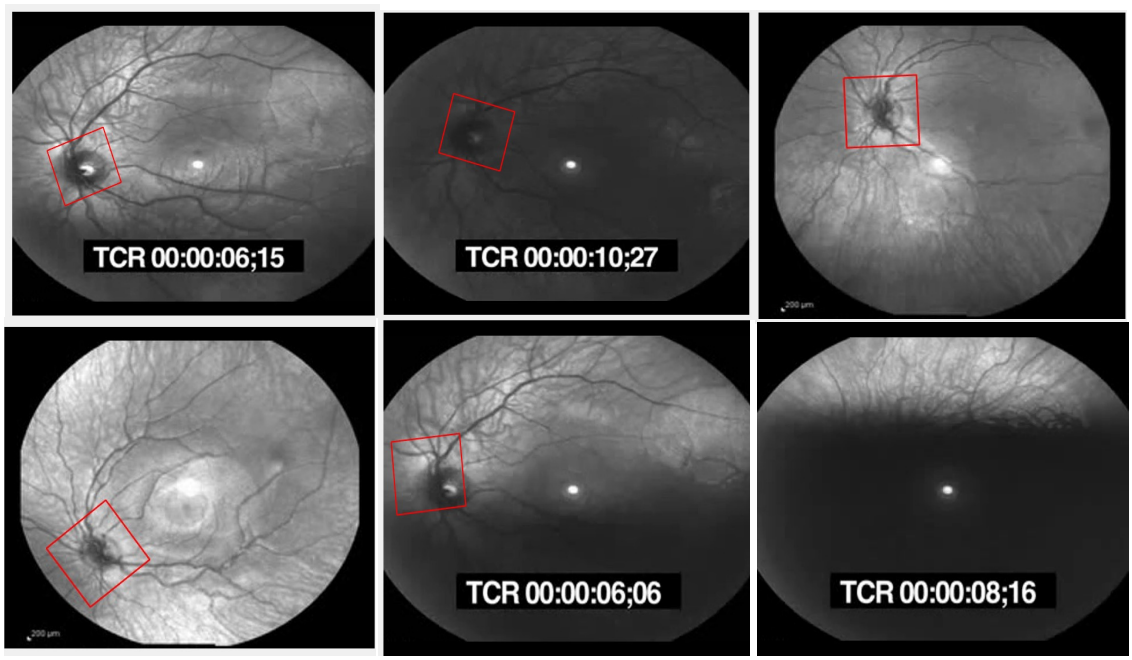
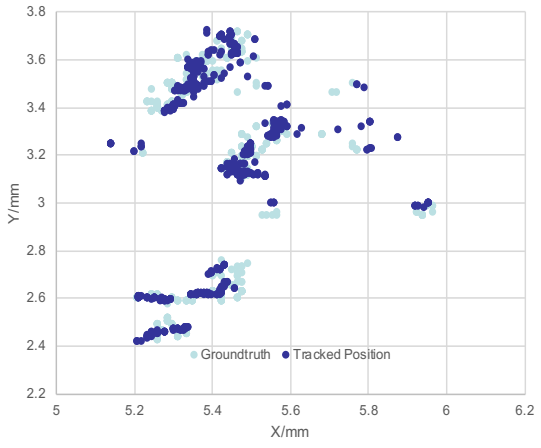
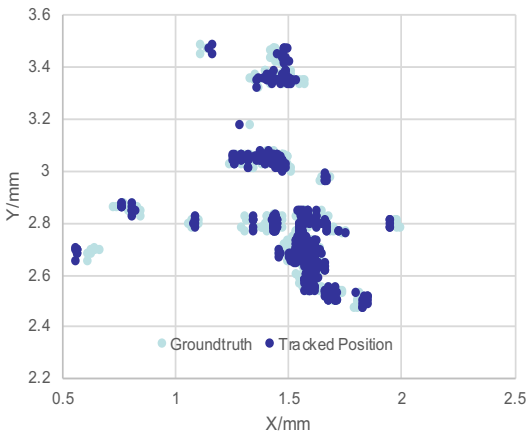


Figure 2.10: Examples of optic disc tracking from two randomly select scanning videos.



(a) Movement positions of optic disc in video 1.



(b) Movement positions of optic disc in video 2.

Figure 2.11: Scatter of the groundtruth and tracked optic disc positions in two videos.

2.2 Intensity-Mosaic: automatic panorama mosaicking of disordered images with insufficient features

2.2.1 Introduction

Image stitching techniques have been well studied in the field of computer vision. Fully automatic panorama stitching techniques from randomly ordered input images are widely used in applications by photographers.[63] Generally, the images they capture with modern digital cameras have high image quality and scenes with well-distributed features. Likewise, in the medical field, it is often necessary to form a composite image from component images for better clinical diagnosis.[64] With the emergence of various low-cost portable imaging devices, the field of view and resolution of captured images are sacrificed for device compactness.[65–67] Moreover, differing from most common images, many biological tissues have a homogeneous background with sparse features. It is also difficult for human users to capture continuous frames from beginning to end with small FOV, and sometimes back and forth scanning is needed to obtain a qualified image. Thus, there is a need for a fully automatic panorama mosaicking method that is robust to a low density of features and can handle disordered images.

The essential step of image stitching is to align input images by estimating a 3×3 camera matrix or homography for each image.[68] There are three main types of image alignment methods in medical image mosaicking: feature-point based methods,[3,67,69] intensity-based methods[65,70] and specific feature or structure matching, e.g. blood vessels and vascular bifurcations.[66,71] Since mosaicking with specific features cannot be generalized to most cases, we discuss more on feature-point methods versus intensity-based (or pixel-based) methods.[64]

Currently most of the mosaicking methods with feature-based registration rely on the high density of features. As surveyed by Szeliski et al.,[72] the common pipeline is to first extract distinctive features from each image, to match these features to establish a global correspondence, and then estimate the homography between the images. Brown and Lowe[63] proposed AutoStitch with scale-invariant feature transform (SIFT) matching under this pipeline, which

becomes one of the most popular mosaicking techniques. The performance of mosaicking is further improved from different perspectives such as seam cutting[73,74]and blending.[75] These approaches can register the images in complex scenarios and are computationally efficient. They assume the feature point pairs are well distributed on the image and can be reliably detected and matched. Although feasible in most cases, the process can fail on low-quality images with sparse features. In the mosaicking of medical images with feature-point based methods,[3,67,69] the input images are assumed continuous, thus it is not necessary to align all images, which makes it easier without considering other homogeneous features from non-adjacent regions.

Intensity-based approaches match the intensity differences of an image pair under a similarity measure, such as sum of squared differences (SSD), Cross-Correlation (CC) and mutual information (MI), then optimize the similarity measure by searching in the transformation space. It makes optimal use of the information available in image alignment, since these methods measure the contribution of every pixel in the image. Thus intensity-based mosaicking is expected to be more robust to images without distinctive features (little texture) than feature-based approaches.[72] In intensity-based approaches, the selection of metric to measure the similarity is important. MI has been compared with CC and found to be advantageous over images with sparse features.[76] SSD is used early in the video mosaicking procedure,[77] then the video mosaicking robustness is improved by using MI.[78–80] In medical field, Hernandez-Mier et al.[81] used SSD for fast endoscopic video stitching. Ben-Hamadou et al.[82] experimentally proved that MI is more robust than the quadratic distance between the grey levels of the two image pixels in terms of scale factor and rotation changes. Miranda-Luna et al.[83,84] also selected MI for the similarity measurement in endoscopic image mosaicking.

However, without the feature matching, establishing a global correspondence with intensity-based methods among disordered input images is problematic. Since the computation of intensity-based registration is generally expensive (slow),[76,82] it is not practical that we register each image with all the other images. On the other hand, for an image set with sparse texture, there will be similar background from non-adjacent images having small intensity

differences, which becomes the interference of the optimization. Thus, the intensity based mosaicking method is generally used in continuous video frames without recognizing overlapped image pairs,[65,70,77–84] or mosaicking tasks with only two to three images.[76,85] This problem can be a restriction in many applications, especially in medical field. Strictly continuous frames with satisfactory quality are challenging to be acquired when handheld imaging devices are used or the imaging tissue is dynamic with interruptions and motions, and sporadic artifacts. Importantly, in cases in which we strive to repeatedly image a region and select the best frames for mosaicking, the input images will be often disordered.

In this paper, we present an intensity-based mosaicking method named Intensity-Mosaic for a series of few-feature images in any order. The schematic of the proposed method is shown in Fig. 2.12. In our work MI is selected as the alignment function in image registration since it outperforms feature-based methods and many other intensity metrics when the image features are not distinctive.[1,64,82] The unique aspect of our approach is that we find the global correspondence of input images in a low dimensional space (LDS) efficiently. It enables MI-based image alignment for disordered data by reducing the repeated MI computation and filtering the interference of the homogeneous background. The image pairs matching and registration is introduced in Sect. 2.2 and 2.3 respectively. Meanwhile, it improves the alignment robustness by only focusing on the registration between nearest image pairs. Since multiple images may overlap in a same region, there can be multiple adjacent neighbors for each image. Selecting the nearest one can provide an initialization for the MI optimization, allowing it to fall into the convergence domain of the MI metric.

The proposed method is validated on three different types of datasets, human retina images captured by a low-cost fundus camera, fluorescence microscopy images of mouse kidney from a hand-held microscope, and bladder phantom images captured with a robotic cystoscope. All image datasets have sparse features and a sequence of images that is not ordered. Our method can generate a complete panorama from disordered inputs with a low alignment error; In contrast, the classical feature-based method AutoStitch fails due to unstable feature detection and matching.

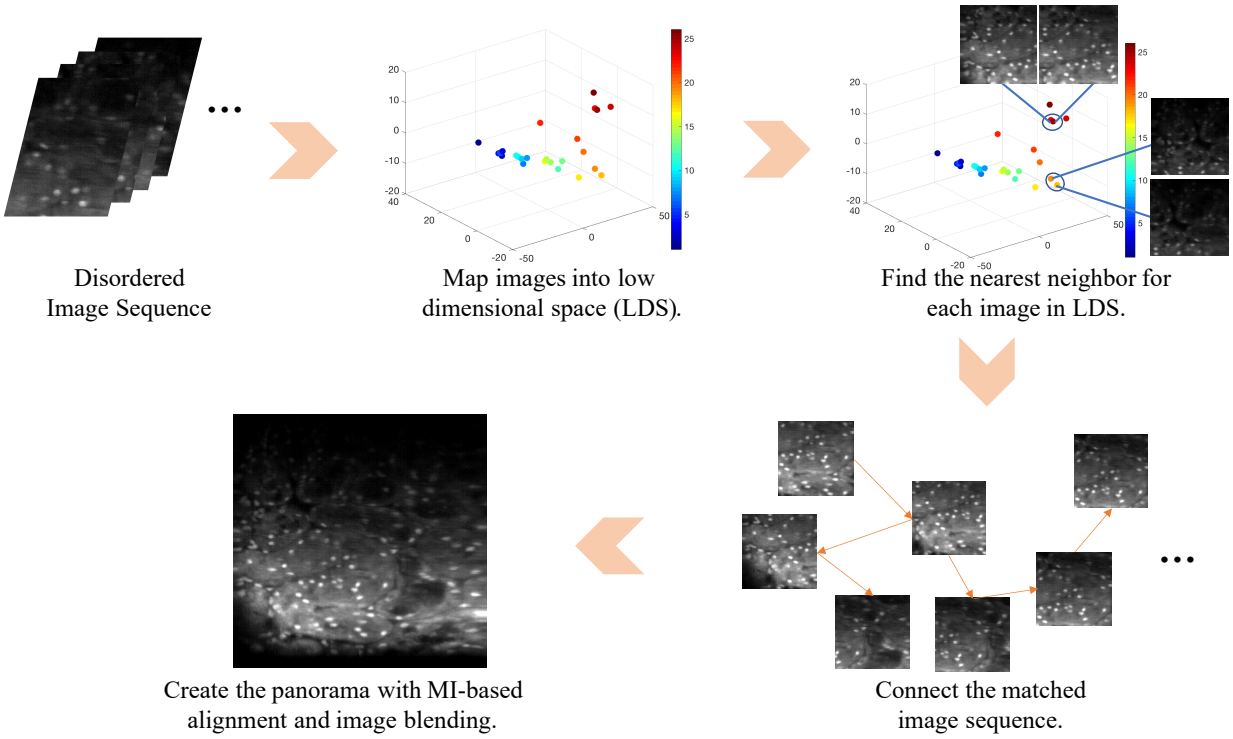


Figure 2.12: Schematic of Intensity-Mosaic.

2.2.2 Proposed Approach

Our method improves both accuracy and efficiency of the image matching part in the common image stitching pipeline for intensity-based stitching. As shown in Fig. 2.12, after acquiring the images in any order, the ensemble of images are mapped into a LDS. We first find the closest overlapped image pair in this space, then each image pair is registered by maximizing the mutual information, and finally connecting the matched images to create the panorama.

Corresponding Image Pairs Detection

To estimate the locational relationship of images in the 2D space, the full image dimension is redundant. Our method uses dimensional reduction to map input images into a LDS, which allows the construction of low-dimensional summaries, while eliminating redundancies and noise in the data. Generally, we can categorize dimension reduction techniques as either linear or nonlinear. The most prominent linear technique is principal component analysis (PCA).[42] Kernel PCA[52] and Isomap[53] are commonly used nonlinear methods. In this work the linear method PCA is selected as the dimension reduction method since it is simple and versatile.

Based on the dimension reduction, the procedure is presented under Algorithm 3 to identify the positional relationship of images to be stitched. For a series of images to be stitched, we resize each input image to a vector $\mathbf{X_i} \in \mathbb{R}^{1\times d}$ and all of them form the matrix $\mathbf{X}$ $\in \mathbb{R}^{n\times d}$, where n is the number of the input images and d is the dimension of the image vector. We obtain the low-dimensional distribution representation of the input image distribution by implementing PCA on $\mathbf{X}$ and compute the mapping $\mathbf{W}$:

$$\mathbf{Z} = \mathbf{X}\mathbf{W}, \tag{2.13}$$

where $\mathbf{W} \in \mathbb{R}^{d\times l}$ maps the image space Ω_1 to a LDS Ω_2 and $\mathbf{Z} = [z_1, z_2, z_3, ..., z_n]^T \in \mathbb{R}^{n\times l}$ is the image representation in Ω_2. l is the dimension of each image representation z_i in the LDS ($l \ll d$). In PCA calculation, $\mathbf{W}$ is the eigenvector of the covariance matrix X^TX. If

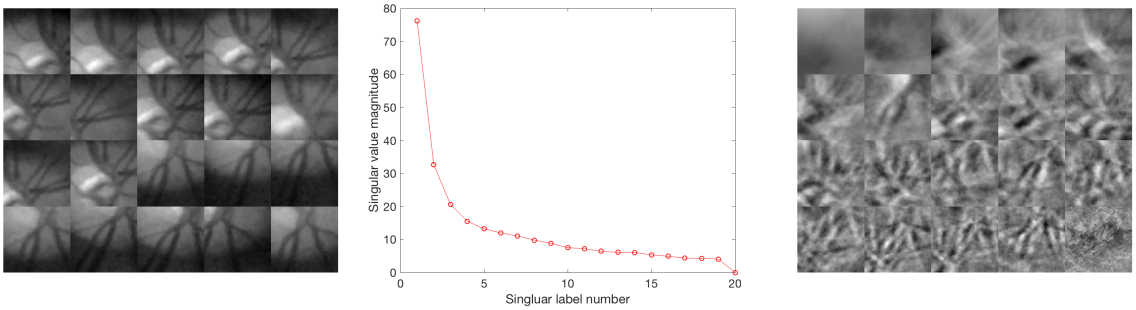
we use SVD to solve the mapping $\mathbf{W}$ in PCA, we do not need to do the eigendecomposition of the covariance matrix $\mathbf{X}^T\mathbf{X}$ but SVD on $\mathbf{X}$ instead. To conduct SVD on a matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, the compute scaling is $d \times n^2$.[86] In PCA computation, tremendous strides have been made to accelerate the SVD and related computations using randomized methods for linear algebra.[87,88] Since we have demonstrated high performance with less than 20 principal components (16 principal components are used in experiments when more than 16 input images), the randomized SVD (rSVD)[39] is used to further improve the computing efficiency, and rSVD has been used in other datasets which are much larger than ours.[38]

Generally, the principal components will focus on the distinctive features on the images such as shapes and contour. Fig. 2.13 presents an example of the principal components of fundus images. Fig. 2.13 (a) shows the 20 input images, and Fig. 2.13 (b) shows the singular values of the input image matrix. We can see the number of dominate components is less than 16. We plot the 20 dimensions of the eigenbases in Fig. 2.13 (c), where the top left one is corresponding to the largest singular value and the bottom right is corresponding to the smallest singular value. With the decrease of the domination, the eigenbases contains more details and noise. In the case of fundus images, the eigenbases concentrate on the vessel structure and the contour of the brightest area in the optic disc.

After the dimension reduction, let $\Delta(z_i, z_j)$ be the Euclidean distance between two image representations in the LDS, which can indicate the distance between images. For each image $\mathbf{X}_i$, we want to find its nearest image $\mathbf{X}_{j^*}$:

$$j^* = \arg\min_j \Delta(z_i, z_j). \tag{2.14}$$

The image pair is then registered with MI-based approach. To improve the algorithm robustness, the top 3-nearest neighbors for each image are first selected for MI registration, and we keep the one with the largest MI value. This chosen method could be optimized for either speed or accuracy by tuning the k nearest neighbors. The details are discussed in Sect. 4. After we matched all of the image pairs, a set of matching images are connected as



(a) (b) (c)

Figure 2.13: Principal components of fundus images in SVD computing: (a) Input images; (b) Singular values; (c) Eigenbases. In (c), the first row shows the eigenbases with the five highest singular values, and the forth row shows the ones with the lowest singular values.

a panoramic sequence.

With the connected sequence, the complexity of following image stitching is $\mathcal{O}(n)$. On the other hand, if we search connected image pairs among disordered inputs with the highest MI value after transformation, the complexity would be $\mathcal{O}(n^2)$ which is much slower than ours. Besides, the homogeneous features can reduce the success rate of global searching. Note that even the efficiency of our method has been improved greatly among intensity-based mosaicking approaches for disordered inputs, the runtime of ours is still longer than feature-point based method in the experiment since MI optimization is more expensive than the feature point extraction and matching.

Image Registration

As described before, the overlapped image pairs $\mathbf{X}_i$ and $\mathbf{X}_j$ are registered with maximization of mutual information.

The main idea of the registration is to find a deformation $\widehat{u}$ at each pixel location x that

Algorithm 3: Image stitching

1 Form the image matrix $\mathbf{X}$ with vectorized input image $\mathbf{X}_i$.
2 Map input images from image space Ω_1 into LDS Ω_2:$\mathbf{Z} = \mathbf{XW}$.
3 For each image $\mathbf{X}_i$:
4 (i). Find the nearest three neighbors $\mathbf{X}_j$ by minimizing the feature distance $\Delta(\mathbf{Z}_i, \mathbf{Z}_j)$.
5 (ii). Registering each $\mathbf{X}_j$ with $\mathbf{X}_i$ with MI and take the adjacent image with highest MI value as the closest image.
6 Find the panoramic sequence as connected set of matching images.
7 Align the sequences with the MI registration results to form panorama $\mathbf{R}$.
8 Panorama blending.
9 **return** panorama $\mathbf{R}$.

maximizes the MI between the deformed image $\mathbf{X}_i(u(x))$ and the image $\mathbf{X}_j(x)$. Accordingly,

$$u_{opt} = \arg\max_{u} \mathrm{MI}(\mathbf{X}_i(u(x)), \mathbf{X}_j(x)), \tag{2.15}$$

where

$$\mathrm{MI}(\mathbf{X}_i, \mathbf{X}_j) = \sum_{i_1 \in X_i} \sum_{i_2 \in X_j} p(i_1, i_2) \log \left(\frac{p(i_1, i_2)}{p(i_1)p(i_2)} \right). \tag{2.16}$$

Here, i_1 and i_2 are the image intensity values in $\mathbf{X}_i(u(x))$ and $\mathbf{X}_j(x)$, respectively, and $p(i_1)$ and $p(i_2)$ are their marginal probability distributions while $p(i_1, i_2)$ is their joint probability distribution. In our work, the transform u for alignment is selected as an affine transformation.

From the MI equation 2.16, we can see the MI function has a discrete formulation which is not differentiable. Therefore several solutions are proposed to smooth the MI function to compute the MI derivatives and maintain accuracy. We use the widely-used method which estimate the joint probability distribution between the images $\mathbf{X}_i$ and $\mathbf{X}_j$ with a Parzen window.[49]

The optimizer used for the MI maximization is based on Newton's method. The MI function is a quasi-concave function, and the parabolic hypothesis of the Newton's method is

only valid near the convergence. When the initial transformation is not in the convex part of the cost function, it will cause the optimization to diverge. Among the image set, one image may have multiple adjacent neighbors with different distances. Generally, not all translation parameters of adjacent neighbors fall into the convex range, thus selecting the nearest image pair in the LDS can improve the robustness and accuracy of the image alignment.

Update of the Panorama

In some cases, the created panorama needs to be updated by new incoming data. For example, there may be absent regions because of disordered imaging, thus more images will be added on the panorama to fill the voids. Besides, when there are slight changes in the scene over time and the panorama is constantly updating itself, it requires a new image to be overlapped onto its corresponding image.

The process for these two conditions are similar, while for filling the empty area, we first need the new incoming images to have overlap with the original panorama so we can build the connection between them. The new images can be mapped into the same LDS to find their nearest neighbors among the original patches, then we can register the image pairs and cover the corresponding region on the original panorama. This process is more like a template matching task where we take the new coming image as the template and the original panorama as the full image. The goal is to match the small FOV template onto the large FOV panorama. More details of the limitation in this template matching case can be found in Experiment A of our previous work.[1]

With a new image, generally we need to match it with every existing image patch to establish the global connection. The computational complexity is $\mathcal{O}(n)$. It is computationally expensive when a large amount of images have been stitched together. Based on the proposed method, the generated panorama can be easily updated with new incoming images without redundant computation, and the complexity is reduced to $\mathcal{O}(1)$. As described in equation 2.18, the low dimensional representation $\mathbf{Z}$ and the mapping $\mathbf{W}$ are computed in the initial image stitching process. We save $\mathbf{W}$ and $\mathbf{Z}$ in the system to keep the low dimen-

sion space. After finding the panoramic sequence, the homography $\mathbf{T}$ between each image and the panorama are stored in a dictionary D. Given a new image $\tilde{\mathbf{X}}$, we can obtain its low-dimension feature $\mathbf{z}_{\tilde{\mathbf{x}}}$ in Ω_2:

$$\mathbf{z}_{\tilde{\mathbf{x}}} = \tilde{\mathbf{x}}\mathbf{W}, \tag{2.17}$$

where $\tilde{\mathbf{x}} \in \mathbb{R}^{\mathbf{1}\times\mathbf{d}}$ is the reshaped vector of $\tilde{\mathbf{X}}$. As described above, the nearest image $\mathbf{X}_j$ can be easily found for $\tilde{\mathbf{X}}$ in LDS Ω_2. We register $\tilde{\mathbf{X}}$ and the top 3 nearest image $\mathbf{X}_j$ to compute their transformation $\tilde{\mathbf{T}}$ and select the one with the highest MI value. According to the known homography $\mathbf{T}_j$ between $\mathbf{X}_j$ and the panorama $\mathbf{R}$, The homography between $\tilde{\mathbf{X}}$ and $\mathbf{R}$ is $\mathbf{T}_j\tilde{\mathbf{T}}$. The new image then can be overlaid on the original panorama without repeated image stitching process.

2.2.3 Experiments

We tested our method on three different kinds of low quality image sets from ongoing research using these medical images: small FOV retinal images, microscopy images of mouse kidney and bladder images captured from a bladder phantom. We include 10 different image sets in each type of the dataset, and the numbers of input images for every image set are 12, 15, 15 for retina, microscope and bladder images, respectively. In Section 3.1, we conduct simulated experiments over a public retina image dataset to test the robustness of our image pair detection under different levels of image degradation. Please note that in the experiments our implemented panorama blending is not optimized, and the blending performance improvement is outside the scope of this work.

We compared our method to classic imaging stitching method AutoStitch,[63] which is one of the most widely used image mosaicking method. AutoStitch uses SIFT features to establish the global correspondence and match image pairs. To show the influence of different parts in our pipeline, we also compare the performance of "dimension reduction (DR) + SIFT" and "Using MI Only" in all experiments. In DR+SIFT we replace the MI-based registration with SIFT feature based registration after matching image pairs. MI-Only does not contain

the image matching with dimension reduction. The way we match images is to optimize the maximum normalized MI (NMI) value between every two images, and select the pairs with the highest NMI as the matched pairs. The matched images are then registered together with the computed transformation in MI optimization. Besides the completeness of panorama, the alignment accuracy is also measured using target registration error (TRE)[57] between the registered image pairs. TRE is a metric to evaluate the image alignment when there is no ground truth of the transformation matrix. The flow chart of TRE calculation is shown in Fig. 2.16. For every two matched image pairs, we take one as the moving image and another one is the fixed image. We need to wrap the moving image to register it with the fixed image together in mosaicking. For every matched image pair, five corresponding landmark points are selected by a trained observer from two images. We let the observer label the same landmarks for three times to compute the observer variability. After the moving image is registered with the fixed image, we can wrap the coordinate of the landmarks on the moving image onto the fixed image. The TRE between two registered images is the root mean square of the distance between the wrapped landmark points and the landmarks on its matched pair. The average TRE of all image pairs is used to evaluate the mosaicking performance.

Validation of the dimension reduction

For the robustness of the dimension reduction part, we conduct experiments to test it under different image degradations over fundus images. We use images from the STARE dataset,[40] which consists of 400 raw fundus images of healthy and diseased retinas. We pre-process each image into size 450×450 pixels. A sliding window is used to crop image patches with size 150×150 pixels from each raw image for mosaicking, and the stride is set to 30 pixels.

Four degradation types in five levels are considered as follows (images are in double format $\in [0, 1]$): affine transform with the rotation/shear parameter of {5°/0.1, 10°/0.2, 15°/0.2, 20°/0.2, 20°/0.3}; additive Gaussian noise with standard deviation varied from 10% to 50% of the pixel value; image blurring with Gaussian kernels with standard deviation of {0.5, 1,

Table 2.6: Success rates of dimension reduction per degradation level.

Distortion level	1	2	3	4	5
Affine deformation	100%	98%	93%	85%	74%
Noise	100%	100%	100%	100%	100%
Blur	100%	100%	100%	100%	100%
Brightness change	100%	100%	100%	98%	95%

1.5, 2, 2.5}; intensity changes of {4%, 8%, 12%, 16%, 20%} of graylevels in the image, which is the nonlinear brightness change.

For each sequence and degradation level, we create 50 image sets for re-stitching the full images as described above. All degradations are applied to the image patches in each input set. To evaluate the performance of the dimension reduction, we take the matched image pairs whose center point distance is less than 90 pixels on the full image as a successful matching. Table. 2.6 shows the success rates of the dimension reduction at every degradation level. The dimension reduction achieves high success rates across the dataset in different degradations, with the exception of two highest levels of the linear deformation.

PCA is optimal for white noise filtering, and the extension is designed for salt & pepper noise, outliers and intensity variation. For instance, robust PCA (RPCA) was introduced to address the issues of outliers, occlusions, and corruption in the data.[89] It decomposes the data matrix into the sum of a matrix containing low-rank coherent structure and a sparse matrix of outliers and corrupt entries. In general, RPCA is more computationally expensive than PCA, requiring an iterative optimization to decompose the original matrix, while it is important to keep RPCA as an option for data with outliers and corruption.

Human Retina Images from a Low-cost Fundus Camera

The retina images are captured with a portable fundus camera called $D - eye$[3] for tele-ophthalmology, which is attached to a smart phone. Fig. 2.14 shows an example of 12

[3]https://www.d-eyecare.com/

disordered input images for mosaicking. Since this mobile device is used outside the clinic and without pupil dilation, the image FOV is small. Image mosaicking is necessary to obtain a more enlarged and diagnostic image of the retina with such devices.[1] When users are imaging the retina with smart phones, although with a high frame rate, a series of disordered quality images of the retina is normal. It is because the handheld device must be aligned with the small pupil by non-expert while the eye is moving with interference from blink and saccade. Thus it is necessary to consider disordered images in retinal mosaicking.

The leftmost columns of Table 2.7 show the mosaicking performance from the retina image datasets. The complete rate is the portion of the sets that a complete panorama can be generated. The TRE values are computed from the complete cases. We can see our method has the highest complete rate of 80% with mean TRE 3.76±1.85 pixels, whereas AutoStitch is 40%. Using DR+SIFT improves the complete rate by 10% in retinal images compared to AutoStitch and remains the same in the following two datasets. For low-cost retina images, the image FOV is very small and only contains a few features. Detecting image pairs with dimension reduction, the SIFT feature matching is restricted within two overlapped images which avoids the interference of other similar regions. However, the main failure reason of AutoStitch and DR+SIFT among these three datasets is for images with non-distinctive features and low image quality. Thus, there are insufficient feature points that can be extracted and matched correctly even within overlapped area, resulting in the failure of image pairs to be registered. This is also the reason we are proposing the mosaicking pipeline for disordered medical images with intensity-based registration methods. Without image pair detection, MI-Only cannot stitch all the input images successfully. The intensity-based methods like MI take all of the pixels into account thus they are easily affected by similar backgrounds from different areas.

Fig. 2.15 shows the example results of the input image set in Fig. 2.14. In this study, we have the same retina imaged by the large FOV fundus cameras at the clinic (Kowa Nonmyd alpha-D III retinal camera) as our baseline. The corresponding region of the mosaicked image is shown in Fig. 2.15(a) which is cropped from the large FOV fundus image and

Table 2.7: Complete rate and TRE comparison over different datasets. TRE and observer variability are in pixels. We mosaic 10 image sets in each type of dataset. The complete rate shows the portion of the image sets that a complete panorama can be generated. TRE is computed from the complete cases and it is omitted when the complete rate is below 20%.

	Fundus images			Microscopy images			Bladder images		
	Complete	TRE	Observer Variability	Complete	TRE	Observer Variability	Complete	TRE	Observer Variability
AutoStitch	40%	3.90±2.03	2.45±1.83	80%	2.98±1.09	0.86±1.21	70%	10.75±19.27	1.56±1.07
DR+SIFT	50%	4.05±2.32		80%	3.23±1.42		70%	9.28±17.84	
MI-Only	0%	NA		40%	1.89±0.83		20%	NA	
Intensity-Mosaic	**80%**	**3.76±1.85**		**100%**	**1.35±0.66**		**100%**	**2.80±4.08**	

provides us a reference for the stitched panorama. Fig. 2.15 (b) and (c) present the stitched panorama of AutoStitch and our proposed method in this example. We can see that our method stitches all of the images and creates the complete panorama, while AutoStitch only matched two images with distinctive features. Fig. 2.15 (d) shows an example of panorama update with new data. It is an example in tele-ophthalmology, where sparse sampling of the retina is taken over time by the user's smart phone device in order to detect disease onset or progression in the retina.[1] In this case, we assume there are pathological changes on the retina, and the yellow and red dots simulate the retinal exudate and hemorrhage. As we introduced before, after we obtain the mosaicked image in Fig. 2.15 (c), the new data can be mapped to the current LDS and connected to the image sequence directly. We can see from the result that our method is robust to such artifacts. For clarity, the new data is highlighted with different color and exaggerated brightness.

Mouse Kidney Images from a Handheld Confocal Microscopy

In vivo real-time microscopy could provide noninvasive pathology data rapidly, which has a number of clinical applications in disease screening and surgical guidance.[90–92] Handheld and endoscopic optical-sectioning microscopes are being developed for noninvasive screening, while most of them suffer from limited fields of view.[64] Using non-visible florescence of labeled disease, it is impossible for users to know if they have adequately imaged the lesion. In this

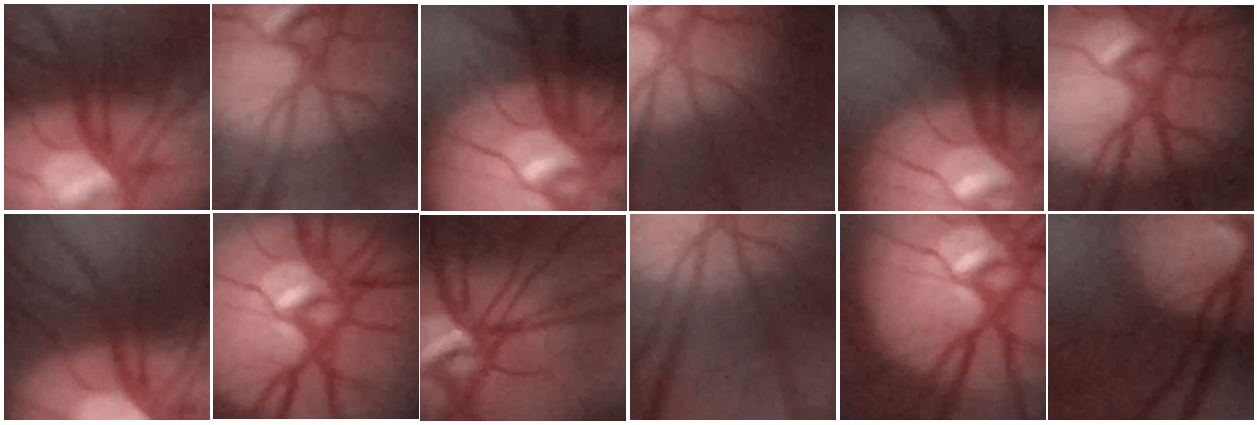


Figure 2.14: Input images from low-cost small FOV fundus camera.

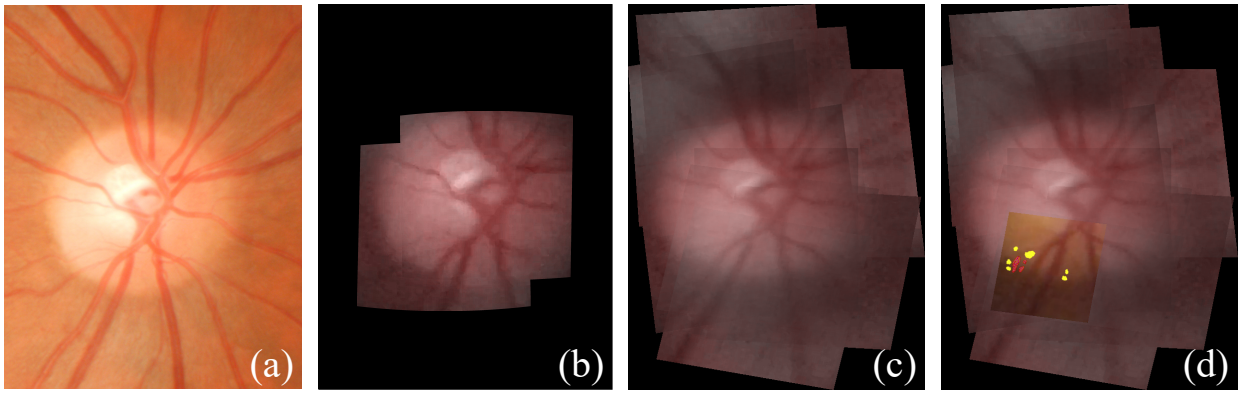


Figure 2.15: Comparison between the AutoStitch and proposed approach: (a) Shows the mosaicked retina region captured by a large FOV fundus camera. (b)(c) are the stitched small FOV images with AutoStitch and our method. (d) shows an example of panorama update with new incoming data with changes. The new data is highlighted in exaggerated brightness and color contrast.

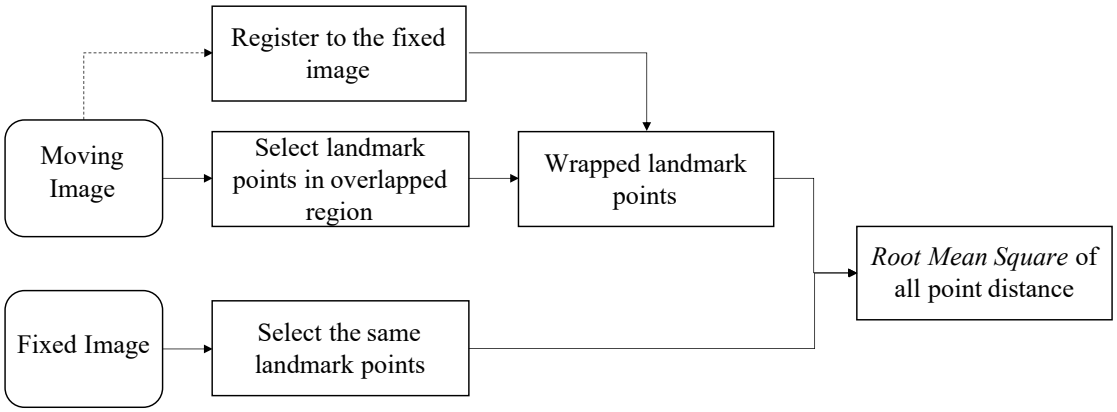


Figure 2.16: Calculation of target registration error (TRE) between aligned images.

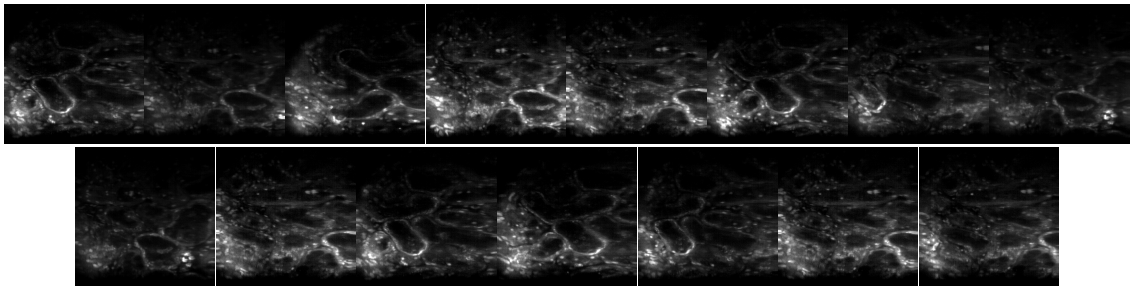


Figure 2.17: Input microscopy images of mouse kidney in one example image set.

experiment, we apply our mosaicking method on images captured by a noninvasive handheld microscope[3,93] for extending overall FOV. This confocal microscope has a visual field of only 350-by-350 μm and outputs images with size of 500×500 pixels. The images are captured on a 3-by-2-mm area of fluorescently labeled (acridine orange) mouse kidney *ex vivo*, from a sub-surface plane that depends on applied pressure on the tissue surface and tissue firmness. We blindly test 10 different sets of the microscopic images, and each group contains 15 image patches. For handheld device that is scanned back and forth to find lesion margins at a depth, the user cannot guarantee all captured images are continuous. Thus our input images in each set are randomly disordered, while assumed to create a continuous mosaic.

The middle columns of Table 2.7 show the comparison of our method and other baselines over the 10 image sets. Similar with the reason in fundus images, sometimes there are insufficient or wrongly matched SIFT features to register the overlapped image pairs, thus AutoStitch and DR+SIFT fails in the same two image sets. MI-Only can complete the panorama in four image sets, whereas our method has a complete rate of 100% by detecting the overlapped image pairs with DR. The average TRE of our method is 1.35 ± 0.66 pixels. Fig. 2.17 shows an input example among the image sets. The features of images are homogeneous with local distortions caused by the bubble deformation, movement during imaging, and slight imaging depth change when hand scanning. Fig. 2.18 shows the stitched panorama of the example set with our method and AutoStitch. For showing the alignment error clearly, the images shown here are without blending. In the yellow boxed area, there are two images containing a bubble when imaging (Fig. 2.18 (a)), while in subsequent images, the bubble becomes distorted and broken (Fig. 2.18 (b)). Influenced by this local deformation, AutoStitch fails to match additional subsequent images at the left end of panorama. In the stitched part of AutoStitch, the seams on the panorama are more obvious than our method. Several alignment examples are boxed in red for comparison.

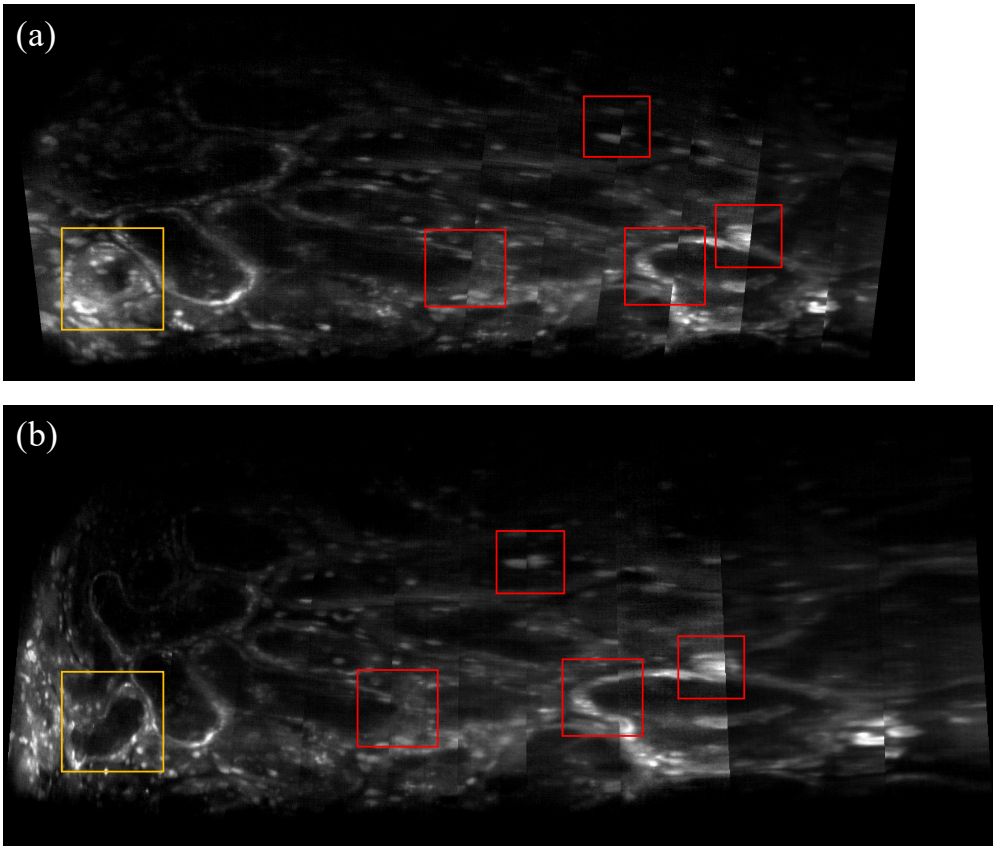


Figure 2.18: Comparison between the AutoStitch and proposed approach over one example microscopic image set: (a) AutoStitch; (b) Proposed method. The yellow boxed region in (a) is a bubble while it is deformed and broken in later captured images as shown in (b). On the left end of panorama, there are two images not being stitched by AutoStitch because of this local change. The red boxes mark several alignment errors in AutoStitch compared to our method.

Video Frames of Bladder phantom for Telecystoscopy

Flexible cystoscopy is an important diagnostic procedure performed by urologists in-office for procedures such as evaluating blood in urine (hematuria), removing stents after kidney stone surgery, and investigating urethral strictures.[94] It is the gold standard for diagnosis and surveillance of bladder cancer, the 6^{th} most common cancer in the US and the costliest.[95] Bladder cancer has a recurrence rate of over 50%,[96] which requires that patients return to their urologists for surveillance cystoscopies up to 4 times per year after initial treatment.[97] Nearly 90% of urologists in the US practice in metropolitan areas.[98] Bladder cancer patients in rural and underserved areas would benefit from a telerobotic cystoscopy system placed in geographically distributed clinics or urgent care facilities, set up and overseen by nurses, and operated by urologists located in their own office.

In this experiment, we stitch captured video frames from a bladder phantom with a robotically actuated flexible cystoscope, as shown in Fig. 2.19. A high-resolution, wide-FOV urothelium[99] was taped to the phantom interior. Cystoscope video was captured during controlled sawtooth-profile trajectories with the FOV of about 115 degrees. The frame rate is near 27Hz and frame resolution is 600×600.

To test the robustness of the mosaicking methods, we captured three different types of videos: (1) The normal videos captured with a trajectory speed of 7 deg/sec; (2) Faster videos with a trajectory speed of 25 deg/sec, which produce more blur on the frames; (3) Videos captured from a phantom attached with a bladder tumor in Ta stage,[100] covering a large region of vessel features. Among the 10 input image sets, four are from video type (1) and three are from type (2) and (3), respectively. We randomly sample 15 overlapped frames from each video to form the input set. Scanning back and forth is often necessary in the cystoscopy because many video frames are discarded during a procedure due to fast tip movement, large distance from surface, bubbles, refractive errors and urine. Thus, video frames used for diagnosis may not be continuous so we disordered the input frames accordingly. The performance comparison with other baselines can be seen in the rightmost columns of Table

2.7. Since the complete rate of MI-Only is very low (20%) in this case, the corresponding TRE is not computed. We can see our method can complete all panoramas with a low TRE of 2.80 ± 4.08 pixels. Two SIFT-based methods have the same complete rate of 70% with high TRE values over 10 pixels. To demonstrate our method can handle larger image set, we show examples in Fig. 2.20 of the generated panorama containing 50 input images in this dataset. The images in the first panorama are extracted from the slow-speed video and the second one is from the phantom with tumor attached. The 50 input images are randomly disordered in both cases.

Since the movement of the cystoscope is in one dimension (see Fig. 2.19), we can control the size of overlap area among input images by changing the sampling interval of the video frames. To ensure the sampling rate is the only variable to control the overlap size, we generate 20 input image sets from the normal videos captured at the slow speed of 7 deg/sec. As tip bending speed increases the overlap area decreases, and we present the effect of the overlap size on dimension reduction and MI-based stitching in Table. 2.8. The DM success rate is the portion of input sets where the correct image sequence can be matched in the low dimensional space. The complete rate is the portion of input sets where the complete panorama can be generated. We can see that DM has a high success rate when the overlap area is larger than 60% of the input images, and drops to 70% with 50% overlap size. One of the main reasons of the success rate drop is the features of the bladder image are very sparse. When the overlap size is reduced to a certain range, the vessel features in the overlap area are less distinctive. The nonlinear deformation caused by the curved bladder phantom surface further reduces the complete rate after adding the linear MI-based registration. With the overlap area larger than 70% in this case, the complete rate of the mosaicking process is over 80%. We expect an improvement of the performance when the input images have less local distortions.

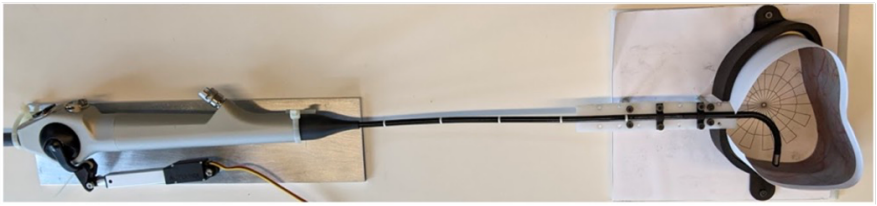


Figure 2.19: Test fixture for robotic cystoscopy data collection with Actuonix linear servo driving tip-bending lever. The 2.5D printed model approximates surface curves that may be seen in cystoscopies.

Table 2.8: Success rate of the proposed mosaicking method under different sizes of overlap area. It is computed over 20 input image sets extracted from bladder videos. The DM success rate is the portion of input sets where the correct image sequence can be matched in the low dimensional space. The complete rate is the portion of input sets where the complete panorama can be generated.

Overlap Area	90%	80%	70%	60%	50%
DM Success Rate	100%	100%	100%	90%	70%
Complete Rate	100%	100%	80%	60%	50%

2.2.4 Discussion and Conclusion

We present an image mosaicking method Intensity-Mosaic for images with insufficient features, using MI-based alignment to stitch multiple images in any order. To make MI-based registration feasible for overlapped image recognition and matching, we apply dimension reduction on the input images and efficiently find the nearest image pairs in a LDS. As demonstrated in experiments, without the image pair detection using DR, MI-Only cannot generate a complete panorama in most cases. To the best of our knowledge, this is the first image mosaicking method that allows automatic stitching of a disordered image sequence with intensity-based alignment. The proposed method outperforms feature-point based method AutoStitch when the image has low quality and sparse features. Based on Table 2, our method compared to AutoStitch exhibits 1.25 (ex vivo microscope dataset) to 2 times (in vivo retina dataset) rate of mosaic completion, and TRE reduction can be 3.8x

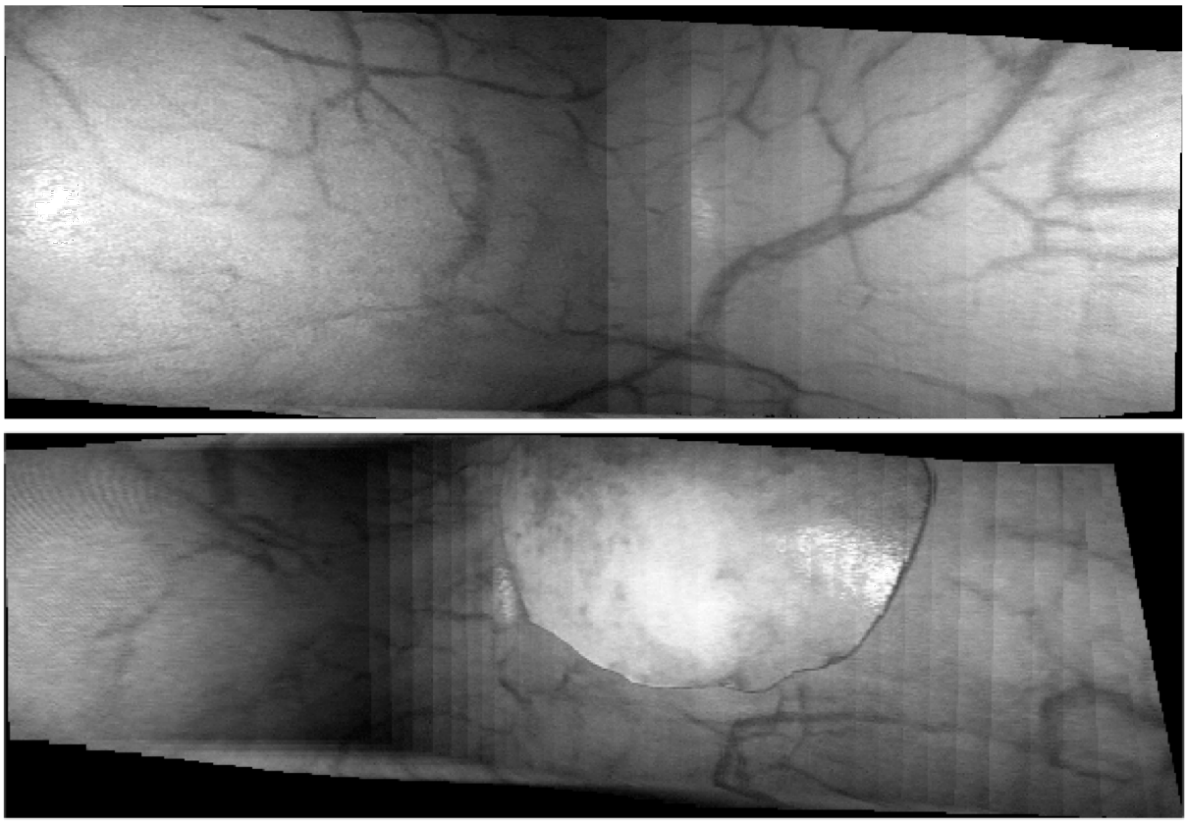


Figure 2.20: Mosaicking examples of 50 input images in the bladder phantom dataset. The top one is from the slow-speed video and the bottom one is from the phantom with tumor attached. Two panoramas are before blending to show the alignment clearly. The robotic cystoscope has vibration in the vertical direction thus the image stitching is not strictly in one direction.

as shown in the bladder phantom dataset.

Intensity-Mosaic is accuracy-oriented instead of speed-oriented. We focus on enabling the intensity-based alignment in disordered image mosaicking when the image texture is sparse and similar. Since the optimization of mutual information between image pairs is more computationally expensive than feature point matching, the runtime of our method is higher than AutoStitch. In the microscope image sets where both AutoStitch and our method can create the full panorama, the average runtime of AutoStitch and ours are 14.09 and 81.98 seconds on CPU. For fair comparison, both of two methods are hybrid programmed with C and matlab. The accuracy of the AutoStitch implementation is similar to the commercial one,[101] which is around five times faster. The proposed algorithm can be further tuned for greater speed while sacrificing some accuracy by reducing the nearest neighbors in Algorithm 1 from three to two or one. If only one nearest neighbor is selected with feature distance, there will be no repeated computation of MI in Step 4 of Algorithm 1 and the speed is around three times faster. Three nearest neighbors are robust in our test data, and this parameter can be changed according to different input images. We believe that our computation speed can be further improved using parallel computing and GPU in the future. Note that in our pipeline, the fine registration part is independent, and the MI-based method can be replaced with other faster intensity-based or even feature-based methods according to the image characteristics. Our framework can also benefit feature-based methods by avoiding the redundant feature matching when sorting and connecting input images: The complexity can be reduced from $\mathcal{O}(nlog(n))$[63] to $\mathcal{O}(n)$.

The fail-safe mechanism is not throughly discussed yet. For the MI registration, there is one failure indicator when the optimization diverges within the given number of the iterations. When the optimization cannot converge the corresponding image will be dropped from the mosaicking, while it does not avoid the convergence to a local optimum. We think the NMI value can also be set as a fail-safe mechanism. When we select the top-3 neighbors with the highest NMI values for each image, we set a threshold of the highest NMI. If the highest MI is less than the threshold, which means the selected nearest neighbors do not have

enough overlapped regions with it, we can drop this patch from the mosaicking pipeline. If a certain number of patches are dropped, we can add human check for our mechanism. If there are too many gaps then we just rerun the pipeline or indicate failure. Other fail-safe mechanisms need further demonstration in the future.

As indicated in the experiments, the clearest potential use of our method is in the medical field. This approach provides a new way to mosaic noncontinuous sparse-feature images and update these over time or to fill in missing pieces immediately after a hand scan by a non-expert. As discussed earlier, our method has the assumption that the rotational variation of input images is less than 20°, which is generally controllable by human users. Currently our experiments are restricted to 2D images, while the technique can be easily migrated to 3D cases, where we want to stitch 3D cubicles into a larger volume. The PCA in dimension reduction part can be done on 3D matrix (principal tensor analysis) to find the adjacent cubicles, then the fine registration is performed.

2.3 Real-time camera localization during robot-assisted telecystoscopy

2.3.1 Introduction

Flexible cystoscopy is an important diagnostic procedure performed by urologists in-office for procedures such as evaluating blood in urine (hematuria), removing stents after kidney stone surgery, and investigating urethral strictures.[94] A diagnostic flexible cystoscopy usually begins with these steps: 1) insertion of the cystoscope into the urethra, 2) inflation and flushing of bladder with clear, sterile fluid pressurized through the working channel (throughout the procedure), 3) inspection of urethral wall during scope insertion, 4) insertion through bladder sphincters, 5) identification of common landmark (usually left or right ureteral orifice), and 6) inspection scan of the entire urothelium (bladder surface) with detailed inspection of areas of interest. Flexible cystoscopy is the gold standard for diagnosis and surveillance of bladder cancer, the 6$^{\text{th}}$ most common and the most costly cancer in the US.[95,102] Bladder cancer has a recurrence rate of over 50%,[96] which requires that patients return to their urologists for

followup cystoscopies up to 4 times per year for surveillance after initial treatment,[97] and a delay in diagnosis of muscle-invasive tumors of 3-6 months can increase risk of death by bladder cancer by 34%.[103] Nearly 90% of urologists in the US practice in metropolitan areas,[98] which can burden some patients with travel costs and time off work.[102] Bladder cancer patients in rural and underserved areas would benefit from a telerobotic cystoscopy system placed in geographically distributed clinics or urgent care facilities, set up and overseen by nurses, and operated by urologists located in their own office. Such a telemedicine system would be useful for many diagnostic urologic procedures, but would be especially useful for bladder cancer patients that require frequent in-person visits for cancer surveillance.

Although this vision of telecystoscopy is not yet in practice, the technologies required have already been demonstrated: the first transcontinental telesurgery was successfully completed two decades ago,[104] telerobotic flexible endoscopes are being introduced commercially for use with surgeons in the room,[105,106] and researchers are developing transurethral surgery robots.[107–109] Introducing teleoperation for bladder inspection is logical because the organ is pliable and not close to critical life-sustaining functions and nurses are well experienced with insertion of urinary catheters. Widespread adoption of clinic-based telecystoscopy will likely begin with a telerobotic platform that can interface with off the shelf, and perhaps single-use cystoscopes[110] which reduces infrastructure overhead. Thus, flexible cystoscopy may serve well as a test case for long-distance teleoperation by urologists in major cities and patients in clinics with nursing and general practitioner support, reducing barriers to timely specialty care.

A major challenge within the teleoperation interface is the accurate pose estimation of the cystoscope within the bladder, since the haptics and proprioception that urologists rely upon for localization will be difficult to simulate in an economical way. Teleoperation of clinical catheter robots has been shown to be improved with the integration of tip-tracking and shape estimation with preoperative 3D anatomical models.[111] Thus, a key feature for developing a telecystoscopy system is the ability to estimate the position and orientation of the cystoscope tip in order to display the pose within a patient-specific model of the bladder and highlight

the current Field Of View (FOV) for the urologist during teleoperation(Fig. 2.21). However, the kinematics of flexible endoscopes can vary widely even between endoscopes of the same make[112] with different amounts of use, and are also dependent on the curvature of the main scope body,[113] making accurate forward kinematics estimation of clinical endoscopes difficult without a detailed characterization for each endoscope. Magnetic field- and electromagnetic wave-based localization strategies are widely used in robotic flexible endoscopy,[114–116] but these methods require extra sensors, specialized hardware, and sensitive calibration. The cost and operational complexity associated with precise endoscope calibration or additional sensing modalities may be disadvantageous to the adoption of a widely distributed telediagnostic platform. On the other hand, an image-based scope localization approach during teleoperation would not only provide the urologist with a feedback of scope pose within the bladder, it could also ensure thorough examination by calculating a running bladder surface coverage metric, providing positions of areas of interest during the current or subsequent procedure, and enabling stabilization around an area of interest.[117]

A standard, image-based approach for camera localization is Simultaneous Localization And Mapping (SLAM). Visual SLAM is common in robotics and utilizes images from monocular, stereo, or RGB+Depth cameras to simultaneously localize robot position and reconstruct the surrounding scene in real time.[118] Visual SLAM has been used primarily in rigid laparoscopic surgery[119–121] and flexible endoscopies.[122] However, the feature detection algorithm and sequential frame matching design in the existing SLAM pipeline does not perform well in many areas of the body due to a lack of texture.[120] Blood vessels on the inner surface of the bladder are a major source of feature points in cystoscope frames, but they are sparse. Structure from Motion (SfM) achieves offline 3D reconstruction through feature detection and matching, triangulation and global optimization of reconstructed 3D points and estimated camera poses, with emphasis on robustness and accuracy, but sacrifices speed. Thus, prior studies used SfM for post-procedure bladder reconstruction.[99,123–127] SIFT is most generally used in SfM because of the high accuracy for feature point extraction and matching,[128] while the computation of SIFT features in SfM is time-consuming. Speeded

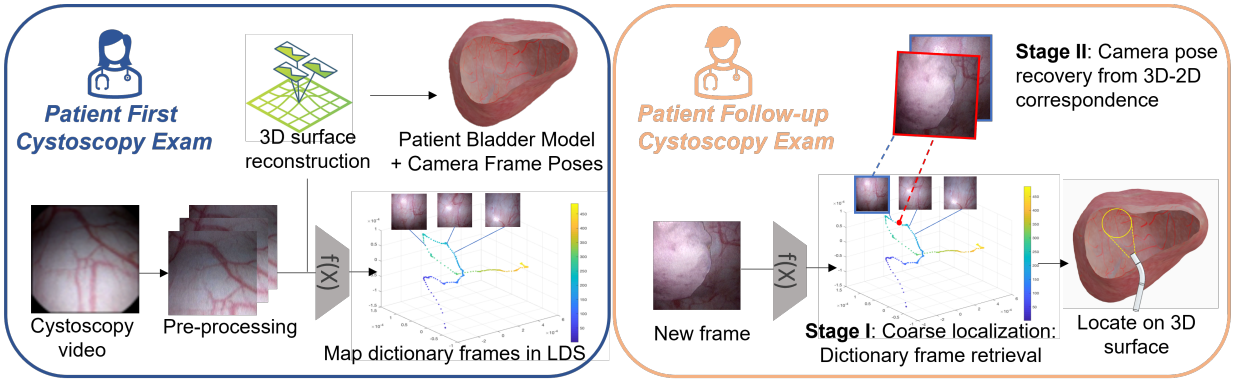


Figure 2.21: Process of our localization system for telecystoscopy. **(Left):** Video from the 1st exam is used to create a 3D bladder model and used image frames are mapped onto a Low Dimensional Space (LDS) as a dictionary set. **(Right):** During the 2nd exam, each new image frame is mapped into the same space and its closest neighbor is retrieved from the dictionary (Stage I). Then 3D-2D correspondences among the new image, its retrieved dictionary image, and the 3D reconstructed model are used to recover camera pose associated with the new image (Stage II). The video frame can then be highlighted on the 3D surface and the estimated cystoscope pose can be used for downstream tasks.

Up Robust Features (SURF) was developed to further reduce the computation load involved in SIFT and provides similar performance at faster speed (˜3×) through the use of integral images.[129] SURF is primarily applied when high-speed matching is required,[130–132] but does not work well under scale or rotation changes, thus, inferior to SIFT for this application. On the other hand, SIFT has limited success with medical images because sparse features and homogeneous backgrounds provide significantly less information for global feature point matching. Low image quality, small FOV, and motion blur in cystoscopy can further increase the difficulty of feature point matching. Accordingly, there will be a high quality requirement of the captured videos for SIFT-based mapping and localization.

In this work, we propose a two-stage global camera localization method for robot-assisted flexible cystoscopy when a video from a previous procedure is available. Recordings of previous procedures could be available for half a million bladder cancer survivors under routine surveillance cystoscopy, which represents a subset of the 724,000 prevalent cases of bladder

cancer in the US.[133] Our method utilizes the initial video for offline generation of a 3D bladder model and a dictionary set composed of frames with calculated camera poses. Then during follow-up exams for surgery or surveillance, our method can estimate camera pose for a new image by first retrieving a prior image with known camera pose and large overlap with the new image frame for coarse localization, and then recovering camera pose from the correspondence information for fine localization in an online manner. Unlike localization based on continuous frames, this coarse-to-fine paradigm performs a global matching, avoiding accumulated errors and the effects of occasional failures. We investigate the performance of our algorithm in localizing video frames and camera pose captured by a servo-actuated cystoscope inserted within a 2.5D bladder phantom and 3D bladder phantom. By changing the settings of scanning and phantoms, we simulate the change of the bladder condition between the first exam and the follow-up exams which may challenge our image-based localization based on a patient's previous exam. For example, we attached the artificial tumors onto the phantom to simulate the tumor progression. We also vary imaging distance to simulate the different extent of bladder distention among different exams. The results showed that our algorithm is reasonably robust to these challenges as well as efficient.

2.3.2 Methods

In this section, we describe the real-time re-localization of the cystoscope camera in the bladder with a prior 3D-reconstructed model generated from the available cystoscope video acquired during screening examination, as in the case of a bladder cancer patient returning for surveillance. In the first visit (Fig. 2.21(Left)), the urologist collects a cystoscope video which fully covers the complete inner surface within the bladder. We first use an off-line 3D reconstruction pipeline[99] to generate a reconstructed 3D model of the bladder inner surface from the video frames. The video frames used for reconstruction are then stored as dictionary set for subsequent re-localization. In the followup visits (Fig. 2.21(Right)), we use the prior 3D-reconstructed surface model as a prior model and estimate the camera pose associated with newly-acquired frame with respect to the coordinate of the prior model.

3D Reconstruction

The shape and texture of the urothelium surface within bladder are reconstructed offline using cystoscope video frames. The 3D reconstruction pipeline is composed of the following modules: 1) *Camera calibration and image preprocessing*: Intrinsic parameters of the cystoscope camera are first calculated from frames imaging a calibration target.[134] Then bladder frames are downsampled to avoid redundancy and preprocessed with adjustment of contrast and illumination as well as correction of lens-induced distortion. 2) *Sparse reconstruction*: An offline SfM algorithm[135] is used to extract and match SIFT features from frames and then calculate the camera pose at each frame as well as a 3D point cloud model depicting the shape of bladder inner surface. 3) *Mesh generation*: Poisson surface reconstruction[136] uses recovered 3D point cloud model to generate a watertight mesh model, which better represents the shape of bladder inner surface. 4) *Texture mapping*: The mesh model surface is then mapped with texture patches cropped from preprocessed frames to generate a textured mesh model,[137] which captures both shape and texture of the bladder inner surface. Thus, the output of the 3D reconstruction includes a textured mesh model that can be used as prior 3D model for the bladder and a dictionary set composed of frames used for 3D reconstruction with their corresponding camera poses, all of which are crucial components for the subsequent camera localization step in followup cystoscopy visits.

Camera localization

Camera localization is a method for computing the camera pose associated with a camera view under a world coordinate system.[138] If we can estimate the camera pose in the coordinate system of the patient's reconstructed 3D bladder model, we can display the real-time location of camera within the model for visualization and also estimate the camera pose under any chosen world coordinate for robot actuation.

To estimate camera pose quickly and accurately, we have developed a novel two-stage camera localization pipeline(Fig. 2.21):

I) *Image retrieval from dictionary set with dimension reduction*: When given a newly-acquired image, we first use an efficient and accurate algorithm to retrieve the nearest dictionary image which has the largest overlap with the new image. This step is a coarse localization of the test frame. The camera pose of the retrieved dictionary frame can be directly used as a fallback solution when speed has higher priority than accuracy.

II) *Camera pose recovery from 3D-2D correspondence*: From Sec. 2.3.2, we already know the correspondence between feature points on each dictionary image and the reconstructed 3D points on the prior 3D model. Thus, we can use the retrieved dictionary image as a bridge to obtain the correspondence between 3D points on the prior model and 2D SIFT features on the new image, in short, 3D-2D correspondence. Then camera pose of the new image can be calculated from the 3D-2D correspondence and represented under the 3D prior map's coordinate system.

Stage I: Image retrieval from dictionary set with dimension reduction

Sampled from continuous video frames during cystoscopy, the dictionary images used for 3D reconstruction have large overlap with their neighbors. Overlap between two images contains correspondence information useful for recovering pose of the camera views associated with the images. Thus with a dictionary set of overlapping images, one can retrieve a dictionary image that has the largest overlap with the newly-acquired image for its pose localization. To perform the retrieval efficiently, we apply dimension reduction and map each dictionary frame into a Low Dimensional Space (LDS) where euclidean distance between frames in the LDS indicates similarity or overlap (*i.e.*, frames that are close to each other in a cystoscopy video are close to each other in the LDS, as shown in Fig. 2.21). Similar to our previous work on retinal images,[1] we achieve dimension reduction by Principal Component Analysis (PCA) through Singular Value Decomposition (SVD), which is simple, versatile, and satisfies the real-time requirement for use in teleoperation, unlike other non-linear methods such as kernel PCA[52] and Isomap.[53] Note that although PCA is known to be sensitive to outliers, occlusions, and corruption in the data, our dictionary images are acquired under expert- or robot-control and selected from the 3D reconstruction pipeline, resulting in good

image quality and minimized number of outlier(bad-quality) images, thus ensuring reasonable performance of PCA.

The procedure of dictionary image retrieval is described as follows. We resize all dictionary images to vectors and form the matrix $\mathbf{X} \in \mathbb{R}^{\mathbf{n}\times\mathbf{d}}$. The low-dimensional distribution representation of the target image distribution is obtained by implementing PCA on $\mathbf{X}$ as shown in Eqn. 2.18.

$$\mathbf{Z} = \mathbf{X}\mathbf{W}, \tag{2.18}$$

where $\mathbf{Z} = [z_1, z_2, z_3, ..., z_N]^T \in \mathbb{R}^{n\times l}$, $\mathbf{W} \in \mathbb{R}^{d\times l}$ and $l \ll d$. The image space Ω_1 is mapped to a low-dimensional space Ω_2 with the mapping $\mathbf{W}$. $\mathbf{Z}$ is the low dimensional representation. We select the top 20 principal components ($l = 20$) to represent each image in low dimension according to the dominant singular values. For more details of the implementation and acceleration, please refer to our previous work.[1]

We define newly-acquired frames from the followup cystoscopy as $\mathbf{T}$, which are represented by the test frames in our experiments. To find the nearest dictionary image to each new frame, we use the same mapping matrix $\mathbf{W}$ to map $\mathbf{T}$ to its low-dimension representation $\mathbf{z_T}$, as shown in Eqn. 2.19.

$$\mathbf{z_T} = \tilde{\mathbf{T}}\mathbf{W}, \tag{2.19}$$

where $\tilde{\mathbf{T}}$ is the vectorized representation of $\mathbf{T}$. Finally, we can quickly find a representation $\mathbf{z}$ with the minimal Euclidean distance to $\mathbf{z_T}$ in the low dimensional space, which corresponds to the dictionary image that has the largest overlap with the new frame.

Stage II: Camera pose recovery from 3D-2D correspondence

To recover the camera pose for the test frame $\mathbf{T}$, we first extract SIFT features $\mathbf{P_T} = \{(u_T^1, v_T^1), (u_T^2, v_T^2), ...(u_T^i, v_T^i), ...\}$ from $\mathbf{T}$, where (u_T^i, v_T^i) denotes the pixel-level position of detected SIFT feature point on $\mathbf{T}$. Then we can match $\mathbf{P_T}$ with the pre-extracted SIFT features $\mathbf{P_D} = \{(u_D^1, v_D^1), (u_D^2, v_D^2), ...(u_D^i, v_D^i), ...\}$ on the retrieved dictionary image. From Sec. 2.3.2, we already know the correspondence between SIFT feature point (u_D^i, v_D^i) and reconstructed 3D point (x^i, y^i, z^i) in the coordinate system of the reconstructed 3D model.

Now using the retrieved dictionary image as a bridge, we can get the 3D-2D correspondence between (u_T^i, v_T^i) and (x^i, y^i, z^i). Each 3D-2D correspondence pair satisfies the projection relation in Eqn. 2.20, where s is a scale coefficient, $\mathbf{K}$ is the camera intrinsic parameter which is known from 3D reconstruction, and the rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times3}$ and translation vector $\mathbf{t} \in \mathbb{R}^{3\times1}$ form the camera extrinsic parameter.

$$s \begin{pmatrix} u_T^i \\ v_T^i \\ 1 \end{pmatrix} = \mathbf{K} \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix} \begin{pmatrix} x^i \\ y^i \\ z^i \\ 1 \end{pmatrix} \tag{2.20}$$

We solve this equation iteratively using Random Sample Consensus (RANSAC) to find the camera extrinsic parameter $\begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix}$. In each iteration, three 3D-2D correspondence pairs are sampled randomly to form an equation group based on the projection relation in Eqn. 2.20. The solution of the equation group $\begin{bmatrix} \tilde{\mathbf{R}} & \tilde{\mathbf{t}} \end{bmatrix}$ are then used to calculate the reprojection error in the test image and count number of inliers based on a chosen threshold. The final $\begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix}$ is selected from the $\begin{bmatrix} \tilde{\mathbf{R}} & \tilde{\mathbf{t}} \end{bmatrix}$ with the maximum number of inliers among all the iterations. Lastly, camera pose can be represented as follows:

$$\begin{bmatrix} \mathbf{R^T} & -\mathbf{R^T t} \\ \mathbf{0^T} & 1 \end{bmatrix} \tag{2.21}$$

which indicates the position and orientation of the camera in the coordinate system of the reconstructed prior 3D model.

2.3.3 Experiments

Robotic cystoscope actuation

We built hardware systems for actuator-controlled cystoscope movement during acquisition of videos, which has been shown to improve extraction of features on bladder phantom in

our prior work.[139]

Hardware setup for 2.5D bladder phantom

A linear actuator was attached to the thumb lever of a Karl Storz (Tuttlingen, Germany) HD-View Flexible Digital Cystoscope (Fig. 2.22(Top)) for servo-controlled angulation of scope tip with desired bending angle. The cystoscope FOV is 100°. The Actuonix (Saanichton, BC, Canada) L12-P Micro Linear Actuator servo is controlled with an Actuonix Linear Actuator Control Board via manual control via potentiometer and digital control from an Arduino Mega. The Control Board provides analog position sensor feedback from the servo. The distal end of the cystoscope is affixed to a raised platform on an independent phantom plate and the cystoscope shaft is kept straight for all experiments.

A 2.5D bladder phantom was made by 3D-printing a bladder-shaped cross section and then taping a high-resolution, wide-FOV panorama of bladder urothelium[99] to the interior surface of the 2.5D model. The bladder contour is designed based on a bladder's sagittal cross-section (Fig. 2.22(Bottom Left)). This 2.5D phantom serves as a simplified test case for evaluating our localization algorithm with limited surface curvature distortions. The size of our phantom's cross section (100x85 mm) is about 3 times larger than that of uninflated adult bladders (83x40 mm on average). The enlarged size guarantees the ideal imaging distance between scope tip and the phantom wall even when the bending angle is large, thus allowing for unconstrained angulation.

To acquire the ground truth angulation for recorded videos, we modeled kinematics for the tip angulation on our specific cystoscope. Prior research on robot-controlled endoscopes[112,113,140,141] describes flexible endoscope angulation in free space as a linear relation between thumb tip and angulation with two additional factors: hysteresis and dead-band. Hysteresis, or backlash, is when the output of a system does not change immediately as the input changes direction. Dead-band is an area around the center of thumb lever travel where angulation does not change. However, this model does not account for curvature of the endoscope body or external contact with the endoscope.

The distal end of the cystoscope was attached below the angulation section and aligned

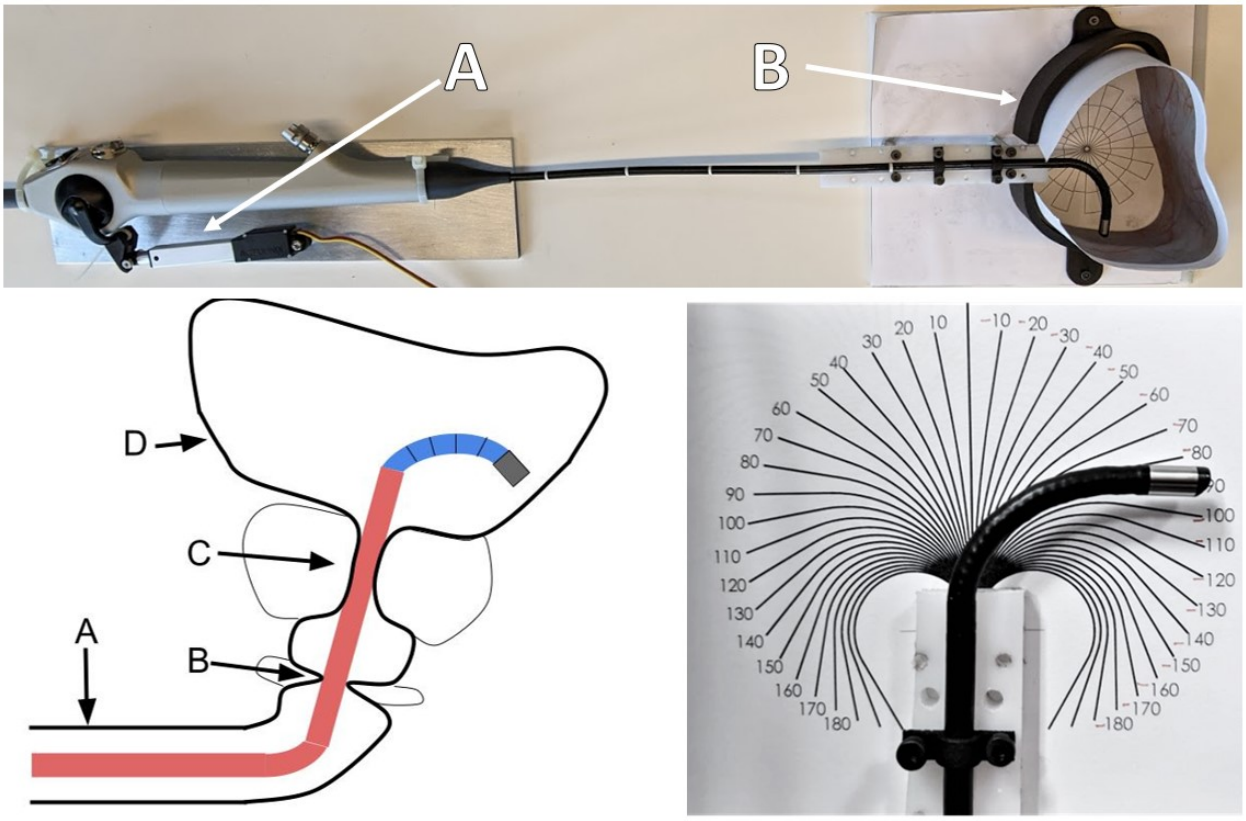


Figure 2.22: **(Top)** 2.5D bladder phantom experiment setup: **A** - linear actuator for cystoscope angulation, **B** - 2.5D bladder phantom. The 2.5D bladder model printed model approximates surface curves that may be seen in cystoscopies. Note that the scope tip is bent with an angulation of 90 degrees in the picture. **(Bottom Left)**: Simplified sketch of cystocope in male anatomy: **A** - Urethra, **B** - External Urethral Sphincter, **C** - Verumontanum at Prostate, **D** - Anterior wall. The flexible cystoscope body is shown in red and controlled angulation area in blue. **(Bottom Right)**: Cystoscope angulation measurement.

over an angulation scale (Fig. 2.22(Bottom Right)). Potentiometer input to the controller was used to step the thumb lever through 3 cycles of angulation. At each step, angulation and linear sensor values were recorded. The linear actuator itself was similarly tested and was found to not exhibit hysteresis. Thumb lever angles were calculated from servo position sensor data.

Hardware setup for 3D bladder phantom

To further study the performance of our camera localization method, we expand the previous 2.5D phantom experiment setup to 3D phantom experiment setup which better simulates the scenario in clinical cystoscopy. A 3-DoF cystoscopy robot (Fig.2.23(Left)) was developed to actuate the same Karl Storz cystoscope and consists of three modules.

A) Flexible cystoscope angulation: The cystoscope's distal section can be deflected from -210° to +140°. The flexible cystoscope shaft is 370 mm long, and the steerable distal section is 60 mm long and 5.5 mm in diameter. A linear servo is used to actuate angulation at the cystoscope's thumb lever.

B) Linear insertion: A ball screw provides the translation action and has a working range of 30 cm. This module consist of a NEMA-17 stepping motor, the ball screw, and a linear bearing, and a slider carriage, which carries the cradle.

C) Cradle with roll module: The cradle for the 3-DoF robot holds the cystoscope and provides rotation along the cystoscope's roll axis. The cradle consists of a 3D-printed body, a small drive pulley linked to a NEMA-17 stepping motor, a driven pulley fixed in a ball bearing, a timing belt, and a mounting point for the angulation servo. A removable clamping ring is mounted on the pulley to fix the cystoscope to the robotic mechanism.

A 3D bladder phantom made by the UW Medicine Center for Research and Education in Simulation Technologies (CREST) is used in the experiments, as shown in Fig. 2.23 (center). The phantom was created by capturing patient data through MRI and CT scanning. The bladder is digitally recognized and isolated by segmentation software and a digital file is created and 3D-printed as a mold. The resulting part represents the bladder volume as a positive form. This form is used as a mandrel to apply layers of platinum-cured and low-

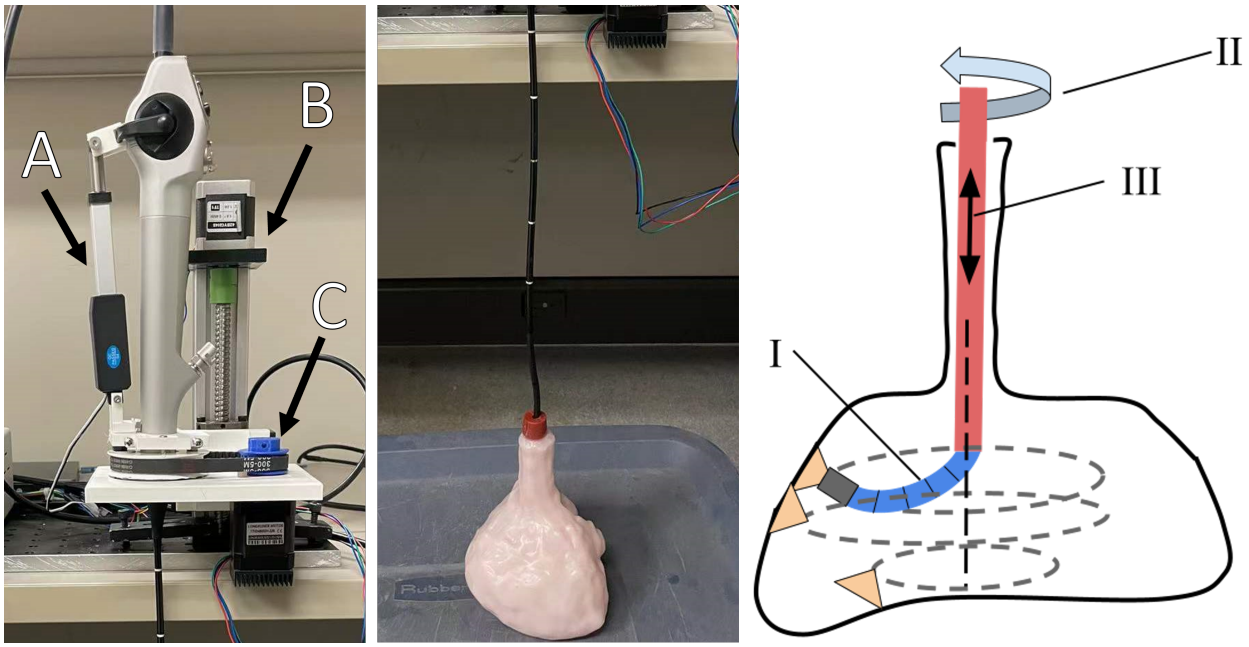


Figure 2.23: 3D bladder phantom experiment setup. **(Left)** The 3 DoF cystoscope robot with three actuation modules: **A** - cystoscope angulation control, **B** - cystoscope insertion control, and **C** - cystoscope roll control. **(Center)** The cystoscope inserted into the 3D bladder phantom. During data collection, the phantom was filled with water and placed in a container among bags of rice to preserve position and shape. **(Right)** Data collection process for 3D phantom. **I** - The bend angle is adjusted to a sufficiently overlapping view (>20%) with the previous scan. **II** - The roll axis is actuated through one revolution clockwise and immediately counterclockwise while a video is recorded. The dashed lines represent the trajectory of the cystoscope tip during video recording. **III** - When the cystoscope hits the walls during a scan, the insertion length is changed and a new set of dictionary and test videos is collected.

durometer silicone material (PlatSil silicone rubber, Polytek Development Corp., Easton, PA) to create the bladder wall. Attention is given to how the layers will be represented by the lighting and imaging from the cystoscope. Many semi-transparent layers are applied to capture depth of the tissue, highlight topology, and represent blood vessels within the phantom. The silicone form is cut and demolded from the mandrel and sealed with adhesive to make the cut line watertight. For simplicity of robot fixation and water filling of the phantom, we kept the 3D phantom inverted during data collection to avoid spilling the water. The robot was fixed on a flat table top above the phantom with some elevation. Although such positioning doesn't influence the performance of our method, in the future we do plan to improve our robot hardware and the sealing accessory of the phantom so that we can distend the bladder to a larger size and manipulate the scope to view the phantom in more optimal perspectives.

Experiments in 2.5D phantom

Dataset

Cystoscope videos were captured with controlled sawtooth-profile trajectories with constant velocity in both directions for 3 cycles, starting and finishing in a downward orientation. Serial data recording included a timestamp in milliseconds, control output, and servo sensor value. The amplitude of the trajectories were set so that the scope tip throughout the trial was not too close to blur bladder features. Trajectory speeds were either slow, medium, or fast (7°/sec, 25°/sec, 60°/sec, respectively). Videos from the cystoscope were saved as MP4s with a framerate near 27Hz and a resolution of 720x720.

We focus on testing the robustness of our method when the video captured in followup cystoscopies differs from the dictionary set in several ways: addition of tumors, distance changes between the cystoscope and bladder surface, and angulation speed changes.

Added Tumors I and II: Considering there may be new tumors emerging on the bladder surface between two clinical exams that may interfere with matching to a prior image, we add tumors to the test videos in two ways: (I) attaching a bladder tumor in Ta stage[100] with tape

when collecting the test videos (Fig. 2.27(1^{st} row)) and (II) digitally adding five different types of tumors in the test video frames (Fig. 2.27(2^{nd} row)). Ta grade papillary tumors were used to test our image matching performance when the original image is obscured with a body of different structure, as would be the case with papillary tumors that grow into the bladder cavity. The tumors were retrieved from online image searches of surveillance cystoscopy and were scaled to our images based on the relative sizes of surrounding vasculature in source images. Digitally placed tumors were added onto test image frames in a random position and rotation.

Imaging Distance Change: During clinical cystoscopy, the bladder is enlarged with water and the enlarged volume may vary between exams by as much as 30%. The inspection distance between the cystoscope tip and the bladder surface will also vary between any two procedures. These factors will cause the imaged area of the same camera location to change between exams, which increases the difficulty of image retrieval from dictionary set (Stage I). We simulate these changes by localizing test frames with a FOV 30% smaller than the dictionary set by translating the cystoscope towards the bladder wall.

Movement Speed Change: During bladder screening and tumor inspection, there may be overly fast movement of the cystoscope leading to motion blur in the frames, which makes traditional tracking difficult. To test our performance during fast movement, we conducted an experiment with test videos with medium and fast movement speed, where the dictionary set is formed from slow speed video.

Evaluation

Quantitative evaluation and comparison with SIFT-only matching: We quantitatively show the performance of the test frame localization by evaluating the success rate of the coarse localization in Stage I and the registration accuracy based on SIFT feature matching between the test frame and the retrieved dictionary frame. The performance of image pair registration is determined by the overlap size and the SIFT feature extraction and matching (influenced by image quality), the former is an indicator of the image retrieval performance and the latter is a crucial step in the 3D-2D correspondence in Stage II of our camera

localization method. Since we do not have a ground truth to directly evaluate the camera pose recovery of test videos yet, we use the registration accuracy to partially evaluate our pipeline. Since cystoscope frames have relatively small size and large inter-frame overlap, they are unlikely to be significantly affected by the nonlinear deformation due to the non-planar bladder surface, which allows for modeling the geometrical relationship between each test frame and its retrieved dictionary frame as a homography transformation for registration.

For comparison, we also present the registration performance with SIFT-only method without our coarse localization. The SIFT-only matching method extracts SIFT feature points from each test video frame to try to match them to features from all of the dictionary images with homography transformation. It takes $\mathcal{O}(n)$ time for each test frame to register with an overlapped dictionary frame globally, where n is the number of dictionary frames. A k-d tree can be used to accelerate the matching process with a time complexity $\mathcal{O}(log(n))$.[63] With the coarse localization in our pipeline, the computation time of the global registration is reduced to $\mathcal{O}(1)$.

We selected 25 test video frames in each test case, sampled randomly and distributed uniformly. To measure the registration accuracy between the test and dictionary frame, we use Target Registration Error (TRE) for comparison. Unlike entropy-based or similarity measures, TRE measures the result intuitively in pixels and is independent of different regularization methods.[1,57] For each test frame, five corresponding landmarks were selected by a trained observer. Two trained observers independently selected the corresponding landmarks from the test frame and the retrieved dictionary image. To obtain TRE for each image pair, we first calculate the homography transformation between the test frame and the retrieved dictionary image from matched SIFT features. Then we use the calculated homography to transform the landmark on retrieved dictionary image to the test frame. Lastly we compute the distance between the transformed landmark points and local landmark points on the test frame. The root mean square of distances for all landmarks and test frames is calculated as the final TRE. A smaller TRE indicates a more accurate homography, which is usually caused by larger overlap and smaller perspective change between the image pair.

Angulation recovery based on image retrieval: To determine the accuracy of our angulation recovery without precisely aligning sensor and video data in each dictionary set, we compare the pixel distance from each test image to its correctly matched image. We then use a linear approximation to determine the tip angulation error from this pixel distance, or the scale between the increment of pixels in localization Δd and angulation $\Delta\alpha$. Taking the arc length as the distance between frames when $\Delta\alpha$ is small (1°), we get a linear relationship between $\Delta\alpha$ and Δd: $\Delta d = K \times \Delta\alpha$, where $K = 20$ pixels per degree.

To fully demonstrate our localization algorithm, we temporally aligned the video frames of the 1st cystoscopy exam video and the medium speed trial with the hysteresis-compensated sensor data corresponding to each video. The test frame's angulation was interpolated between the angles associated with its two nearest dictionary images. Since there is large overlap between close dictionary frames, we assume linear movement between continuous dictionary frames and interpolate accordingly.

Experiments in 3D phantom

Dataset

The same Karl Storz cystoscope was inserted into the water-filled 3D bladder phantom during the experiment, thus camera instrinsics parameters and other camera-related parameters are assumed to be unchanged from those in the 2.5D phantom data, except camera trajectory. Fig. 2.30 shows several examples of the video frames collected from the inner surface of the 3D bladder phantom. The vessel features are much denser and thicker than the printed clinical bladder images shown in Fig. 2.27, and extra features are included, such as fixed bubbles, seams, shadows from surface topology, and floating particles.

Scanning of the 3D bladder phantom was performed in a series of circle trajectories enabled by rotating the cystoscope along its roll axis. Each circle trajectory has a fixed bend angle and the bend angles of different circles increases in the series, as sketched in Fig.2.23(Right). All scanning is performed in a slow and constant moving speed of one circle per minute. The cystoscope was only able to image about half of the bladder surface with

this simple trajectory before the distance from the bladder wall became too small. Full imaging of the bladder during cystoscopies requires larger distension of the bladder through pressurized fluid filling and precise, coordinated actuation of the cystoscope with respect to the anatomy that our current robotic platform is not yet capable of.

To further test the robustness of our method in the 3D phantom, two parameters are varied during data collection for two groups of experiments.

Tip Bending Angle Change: The first group of experiments aim to evaluate the performance when there is limited overlap between the dictionary images and the test images. Since our scanning is performed layer by layer, we control the view overlap by changing the bending angle of the cystoscope tip. Test scans are recorded at bending angles between those of the dictionary scans at the same insertion depth within the bladder. The test images have 10%-25% vertical shifting with the dictionary images, and they are divided into levels of tip bending I and II. Note that these test videos still contain perspective changes and other potential local deformations because they are separate scans.

Insertion Depth Change: The second group of experiments aim to evaluate performance with changes in the imaging distance during cystoscope scanning which simulates the bladder volume variation between different exams. We set different insertion depths of the cystoscope to change the distance during the test video scanning. Three different insertion levels I, II, III are used which are 2.5mm, 5mm and 10mm from the insertion depth used in the dictionary video. With the insertion depth change, there is also trajectory shifting between the test and dictionary scanning.

Evaluation

Similar with the 2.5D phantom case, we also compare our method with SIFT-only matching by the success rate and mean TRE. Within each level of changed tip bending angle and insertion depth, 100 test frames are sampled and coarse-localized with the dictionary set.

Due to lack of reliable ground truth for camera poses, our camera pose recovery is qualitatively demonstrated. We visualize the trajectory of recovered camera poses (both translation and orientation) for the test video frames in tip bending angle II with respect to the recon-

structed 3D model. Since the test videos are acquired by scanning the bladder phantom in circles as in Fig. 2.23 (Right), we can visually evaluate the quality of recovered camera poses.

2.3.4 Results

Hysteresis model in 2.5D setup

Tip angulation kinematic data (Fig. 2.24) shows a hysteresis of 6.5° at the thumb lever. No discernible dead-band is observed. The resulting parallel hysteresis model is: $\alpha = 5 \times \theta \pm 16$, where α is the angulation angle in degrees and θ is the thumb lever angle difference from center. The angulation estimation looks for inflection points, at which it maintains its estimate and either: switches the model when the thumb lever has continued in the new direction past the 6.5° horizontal gap; or returns to the original model when the lever angle movement matches the original direction and passes the initial inflection point. The 32° vertical gap between models at a given thumb lever angle represents the imperfect precision of kinematic estimation when the direction of the thumb lever movement is unknown or not modeled.

3D Reconstruction

To evaluate the accuracy of reconstruction, we first align the ground truth model of the phantom and the reconstructed model in Meshlab.[142] We then use Meshlab to calculate Hausdorff distance, which represents the upper bound of accuracy of all reconstructed points.

Reconstruction of 2.5D phantom

Using 156 frames as input, the offline 3D reconstruction takes 560 seconds (9.3 min) on average. After the bladder phantom reconstruction is aligned with ground truth as in Fig. 2.25, the Hausdorff distance is calculated to be is 0.0319 (normalized over diagonal of bounding box), *i.e.* the error is bounded within 3.2% of the size of the phantom. Refer to our prior work[127] for detailed instructions on model alignment for evaluation of shape

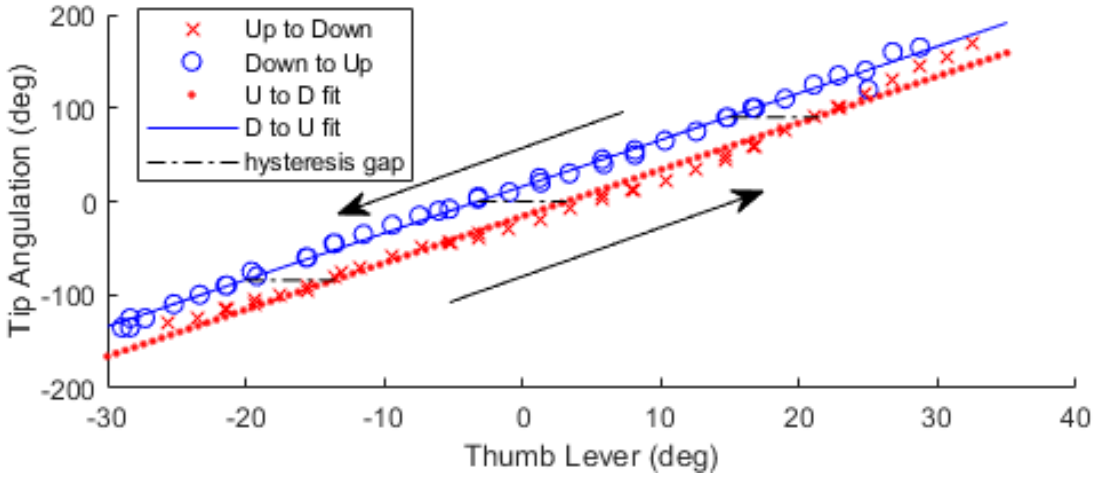


Figure 2.24: Hysteresis model of cystoscope angulation. When the direction of the cystoscope changes, the estimated value is held constant until the sensor value returns to the point of change or crosses the "hysteresis gap", the horizontal distance between the parallel lines.

reconstruction.

Reconstruction of 3D phantom

Using 548 frames as input, the offline 3D reconstruction of the 3D phantom takes 1928 seconds (32 minutes) on average. After the bladder phantom reconstruction is aligned with ground truth, the Hausdorff distance is calculated to be is 0.0290 (normalized over diagonal of bounding box), *i.e.* error is bounded within 3% of the size of the phantom.

Note that the ground truth shape of 3D phantom is acquired from a 3D scan of the mold that was used to make the 3D phantom. And since the phantom slightly expands when filled with water, it is expected that the reconstructed surface model is actually larger than the original model, as shown in Fig. 2.26. This means that the reconstruction may have better accuracy than what is shown by the calculated Hausdorff distance.

Camera localization

Localization results on 2.5D phantom

The performance of our localization approach and SIFT-only approach among sampled

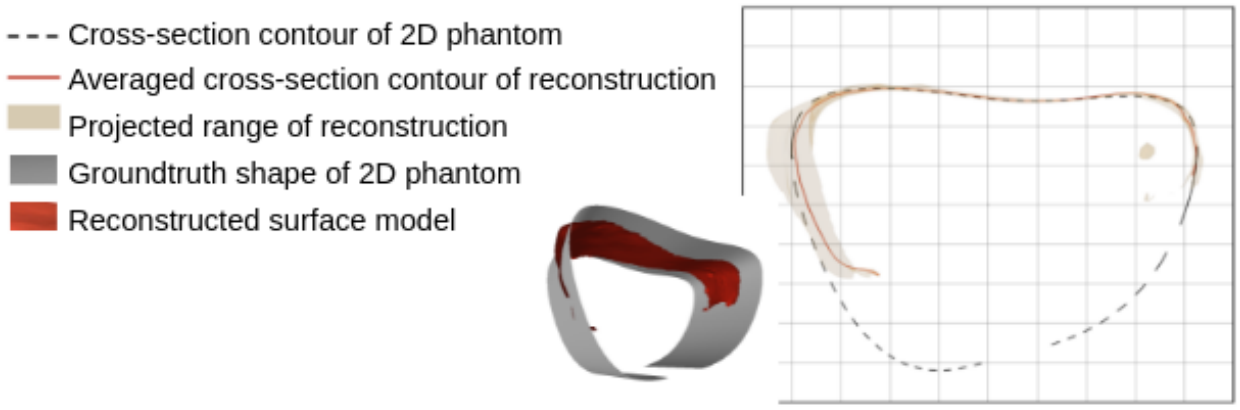


Figure 2.25: Comparison between reconstructed model of cystoscope scanned surface and the ground truth shape of 2.5D bladder phantom. Inset plot shows the reconstructed surface model (red) aligned with the 3D surface ground truth surface shape of the phantom (gray).

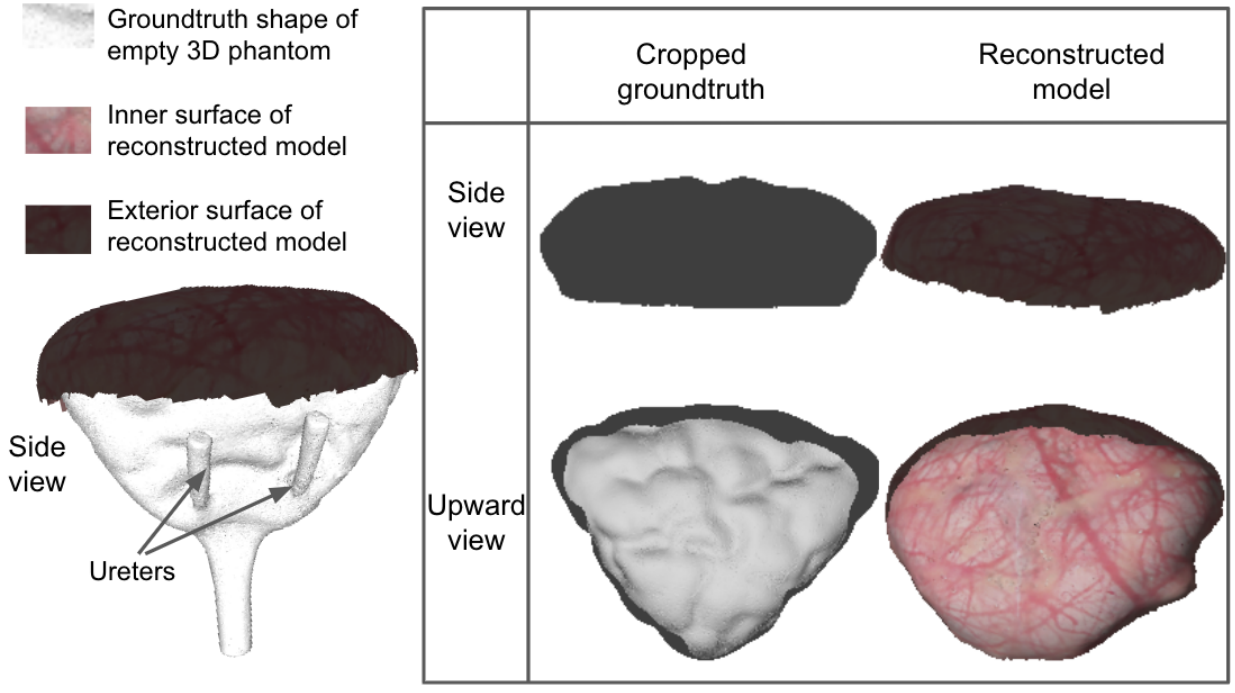


Figure 2.26: **(Left)**: Comparison between reconstructed model of cystoscope scanned surface and the ground truth shape of 3D bladder phantom. **(Right)**: Side view and upward view of cropped ground truth shape and reconstructed surface model of 3D bladder phantom.

test frames is defined by success rate, runtime and average TRE of successful matches, as shown in Table 2.9. The success rate of the SIFT-only control method is defined as the percentage of successful matching pairs with TRE less than 15 pixels. The success rate of Stage I in our method is defined as the percentage of test frames matched with a correct dictionary image with recognizable overlap. We also perform SIFT-based fine registration (denoted as Reg. in table 2.9) on the test frame and its retrieved dictionary image to calculate TRE. Thus success rate of our method followed by the fine registration is also calculated as for the SIFT-only method.

Since we also match SIFT features in registration, the TRE of SIFT-only among successful cases are similar with ours over the 2.5D phantom. In every experiment our method has a significant improvement over SIFT-only in success rate. Except when changing FOV via imaging distance, our success rate is over 96%. Our method can reach an accuracy of less than 10-pixel TRE with an average observer variability of 2.98 ± 1.64. When the FOV changes, the success rate of registration is 80%, while most test images can be matched with a correct dictionary image in Stage I (96% success rate).

The coarse localization by Stage I of our method improves the running speed of fine registration of test images to a correct match among hundreds of dictionary images to around 20 times faster than using a SIFT-only global registration method. Matching the test frames in the LDS happens at about $\sim 20 fps$, which is over 100x faster than SIFT-only. Several success and failure examples with challenging test video conditions are shown in Fig. 2.27. Failure cases of the subsequent fine registration indicate there are insufficient matching SIFT points after the correct image retrieval.

In the coarse localization, we select the top 20 PCA coefficients to form the low-dimension representation of the dictionary frames in LDS. Fig. 2.28 shows a distance map calculated from 200 dictionary frames (continuously sampled from a video sequence) after they are mapped into the LDS. In the distance map, the intensity of a square at row i and column j indicates distance between low-dimension representations of frame i and frame j in LDS. We can see the shortest distance values is gathered near the diagonal axis. It confirms that

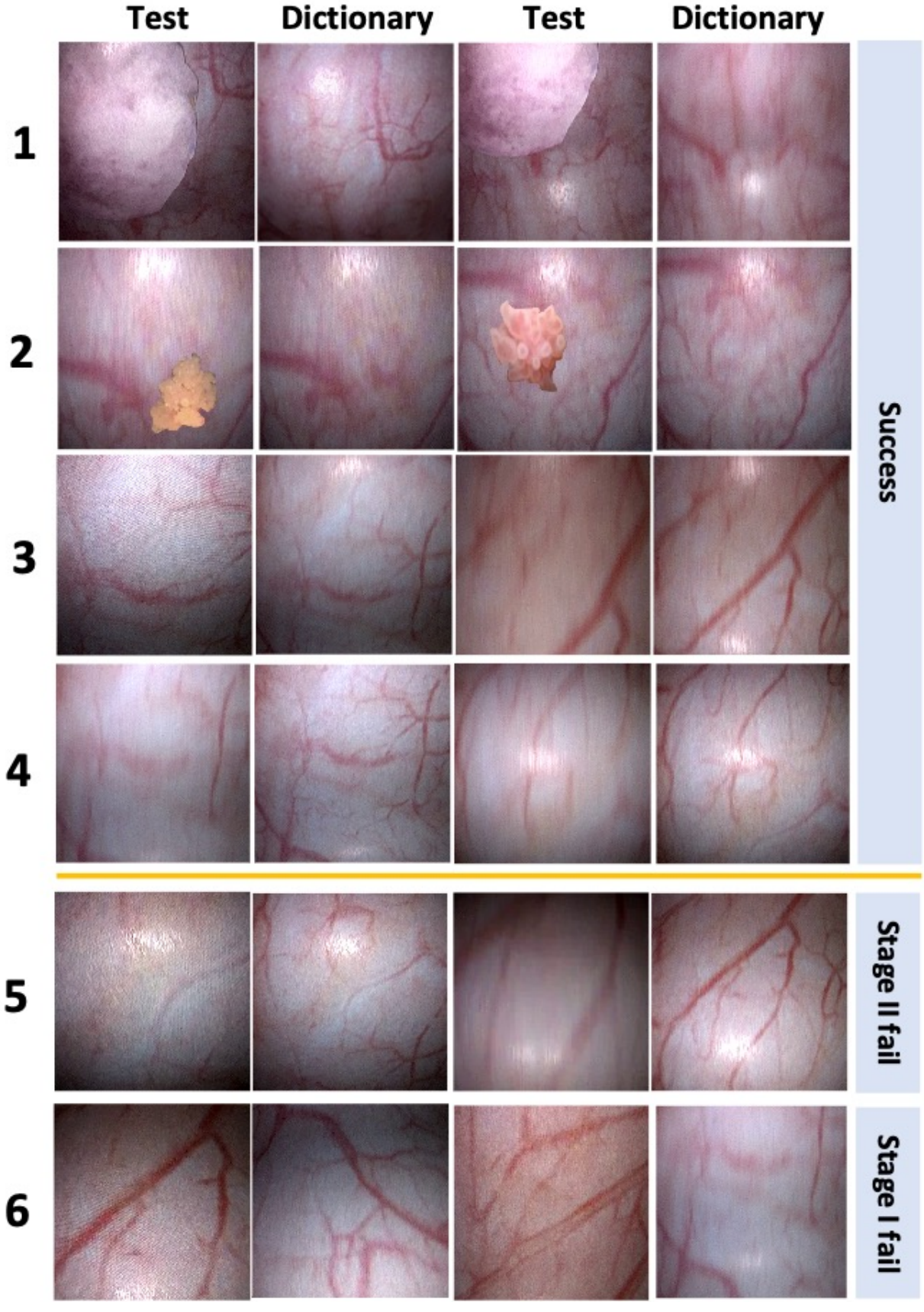

Test
Dictionary
Test
Dictionary
1
2
3
4
5
6
Success
Stage II fail
Stage I fail

Table 2.9: LOCALIZATION PERFORMANCE PER IMAGE FRAME OVER 2.5D PHANTOM

	SIFT-only		Ours		
Changes From Dictionary	**Success Rate**	**Runtime**	**Success Rate (Stage I/Reg.)**	**Average TRE (Pix)**	**Runtime (Stage I/Reg.)**
Tumor I	80%	6263ms	100 / 96%	5.6	53 / 368ms
Tumor II	84%	6482ms	100 / 100%	3.43	42 / 332ms
Distance	56%	6977ms	96 / 80%	8.4	51 / 291ms
Speed (Med)	72%	6087ms	100 / 100%	6.7	55 / 304ms
Speed (Fast)	76%	6115ms	100 / 100%	5.9	47 / 312ms

Table 2.10: ANGULATION PREDICTION SUCCESS RATE

Tumor I	**Tumor II**	**FOV change**	**Speed (Med)**	**Speed (Fast)**
96.5%	98.2%	85.1%	99.7%	97.3%

in the LDS, each image is still closest to its adjacent frames, which should have the largest overlap with the image.

Table 2.10 shows the percentage of robot angles computed from frame localization with an error less than 5° compared to the pixel-based linearization of the averaged robot trajectory (*i.e.*, within 100 pixels of the sawtooth trajectory in each trial). Each test video is downsampled to 370 frames. We only take the results in Stage I when the subsequent fine registration fails. In most cases, the robot tip position can be correctly estimated with a success rate over 96%. Changing the test video FOV (imaging distance) by about 30% increases the difficulty and the success rate is only 85.1%. The angulation trajectory reconstructed from the medium speed trial is seen in Fig. 2.29. The RMS trajectory error over the 23 second long trial is 9.4° and the RMS error between the coarse and fine estimates is only 1.3°.

Localization results on 3D phantom

Table 2.11 shows the the success rate, runtime and the average TRE of successful matches

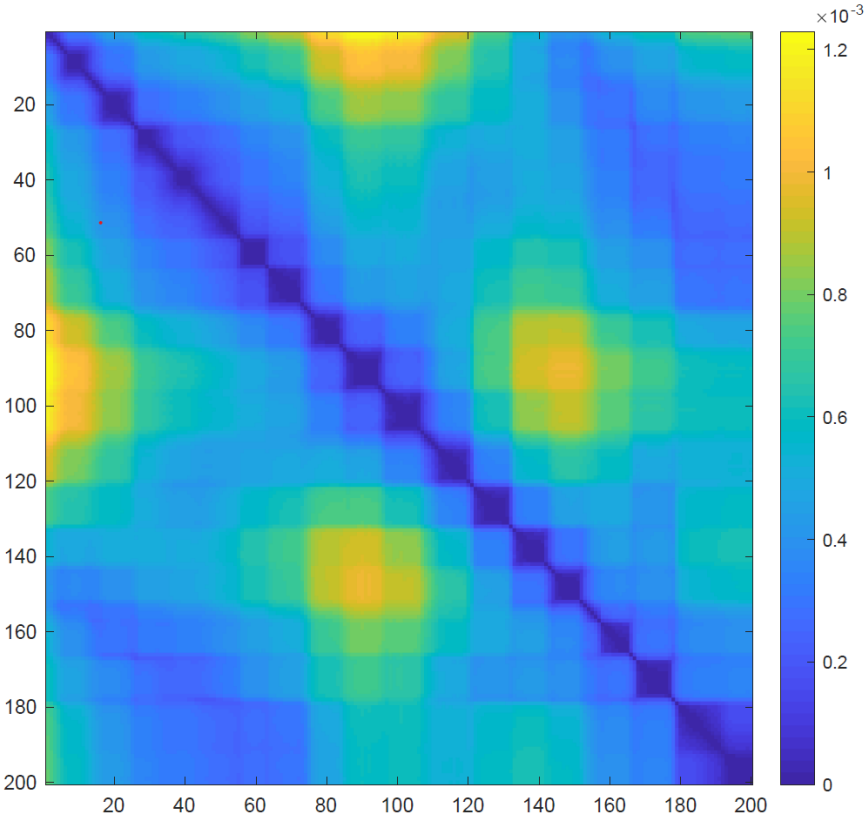


Figure 2.28: Distance map of 200 dictionary frames in LDS. Dark values indicate small distances in LDS and larger overlap between the image pair; light values indicate large distances in LDS and smaller or no overlap between the image pair.

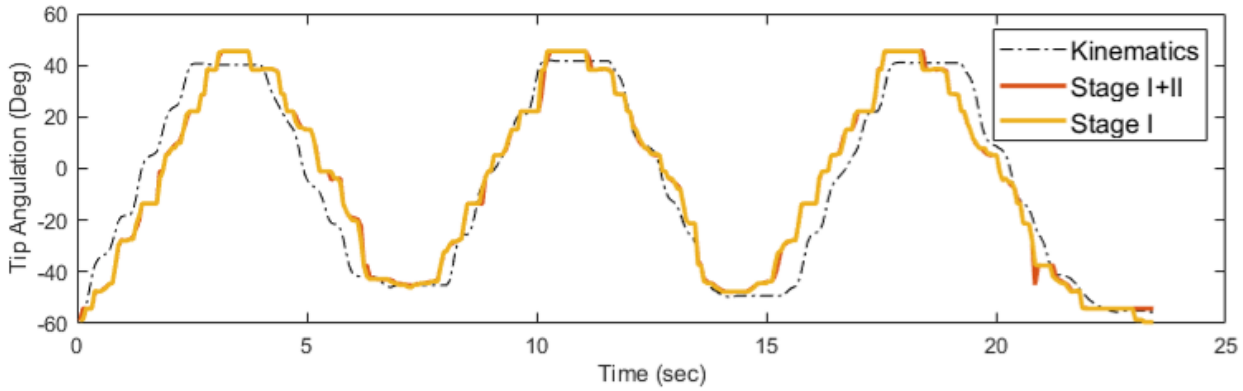


Figure 2.29: Angulation trajectory computed from our localization method during the medium speed trial. The dictionary images are paired with kinematics estimates synchronized with video recording and the localized trajectory is compared to kinematics estimates from the same trial.

of our coarse localization + fine registration approach and SIFT-only approach among different test videos. The success rate and TRE are defined to be the same as in the 2.5D phantom case. Except for the Insertion Depth III test, our success rate is over 99% in all cases. Our method reaches an accuracy of less than 3-pixel TRE with an average observer variability of 1.32 ± 1.02. With sufficient distinctive feature points, SIFT-only method in these experiments has a high success rate, however, it is very time consuming with a runtime of each test frame around 60-75 times slower than our method. The coarse localization (Stage I) is over 1000x faster than SIFT-only method. In the case of insertion depth III, our success rate is 4% lower than the SIFT-only method. The SIFT-only method can sometimes find the correct match with the overlap of selected matched pairs less than ours, especially in insertion depth change, thus we have a smaller TRE among success matches in these cases. Several success and failure examples under different types of test videos are shown in Fig. 2.30.

Figure 2.31 visualizes an example of the camera localization results. In this example, the dictionary images are acquired in three circles with different tip bending angles and the same insertion length to achieve 3D reconstruction. Fig 2.31(Left) shows the camera poses

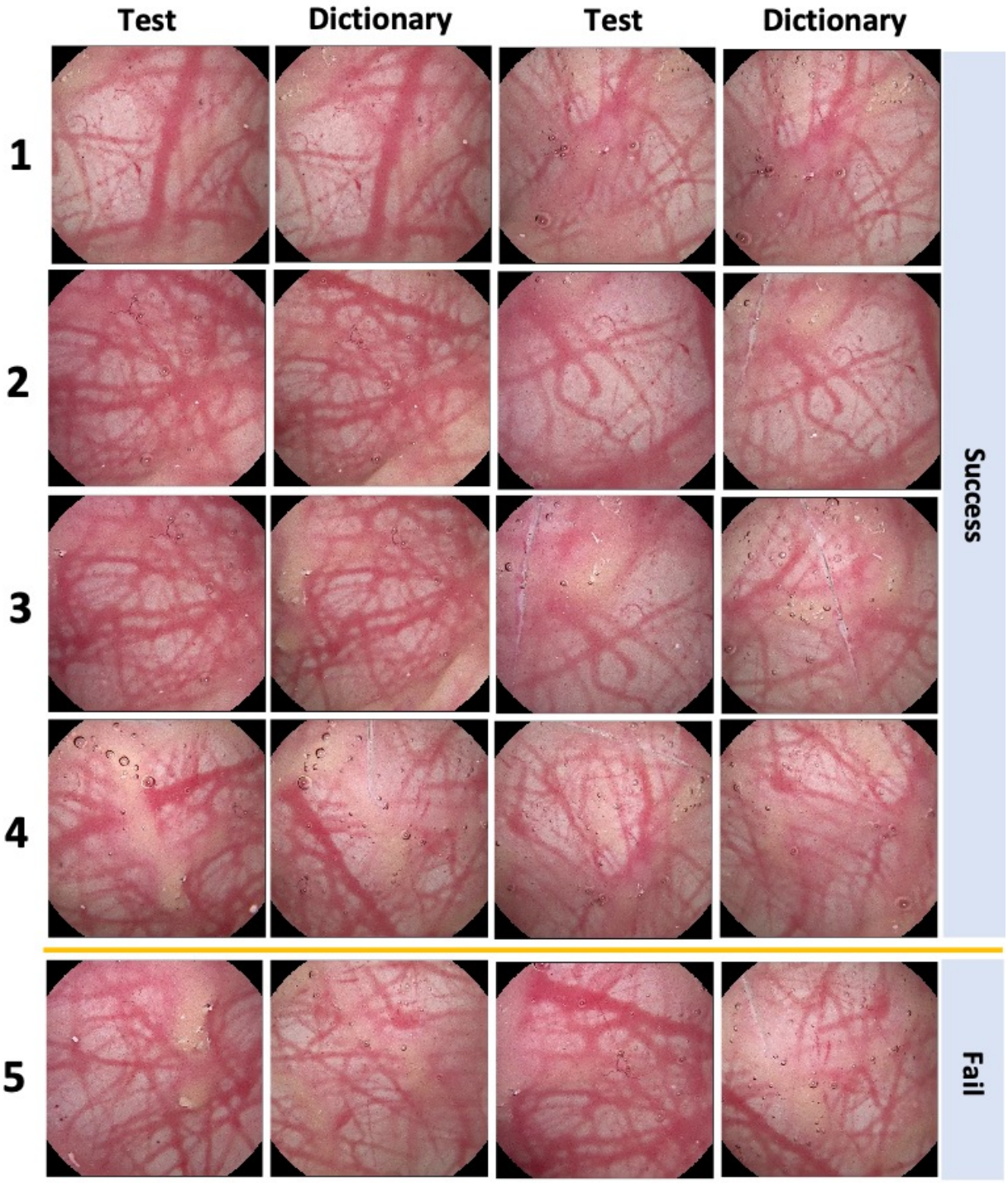


Figure 2.30: Test frames and retrieved dictionary images of success and failure examples of our algorithm within the 3D phantom. *Row 1*: success examples in tip bending angle change; *Row 2-4*: success examples in insertion depth change; *Row 5*: Failure cases in insertion depth change.

Table 2.11: LOCALIZATION PERFORMANCE PER IMAGE FRAME OVER 3D PHANTOM

	SIFT-only			Ours		
Changes From Dictionary	**Success Rate**	**Average TRE(Pix)**	**Runtime**	**Success Rate (Stage I/Reg.)**	**Average TRE(Pix)**	**Runtime (Stage I/Reg.)**
Tip bending I	100%	1.86	38676ms	100%/100%	1.81	43ms/602ms
Tip bending II	100%	2.53	37123ms	99%/99%	2.20	41ms/619ms
Insertion I	100%	2.56	38965ms	100%/100%	2.37	46ms/634ms
Insertion II	99%	5.09	39012ms	99%/99%	2.82	43ms/645ms
Insertion III	98%	5.12	37841ms	94%/94%	1.98	42ms/622ms

(denoted by solid red frustums) of all dictionary images the point cloud of the reconstructed 3D model (denoted by blacks points). The test video in tip bending angle II has a tip bending angle between the top two largest angles used in the dictionary set. With the two-stage camera localization pipeline, we found the subset of 3D points from the reconstructed 3D point cloud that are visible in test frames. This subset appears to be a ring (Fig. 2.31(Right)). We then extracted 3D-2D correspondence based on the matching relation among test image, its corresponding retrieved dictionary image and the reconstructed 3D point cloud. And finally the camera poses are recovered as shown in Fig. 2.31(Right), which appears to be a circle trajectory with camera facing towards the phantom wall. There is only one outlier below the point cloud whose recovered camera pose is clearly wrong.

2.3.5 Discussion

Our two-stage camera localization method can provide pixel-level accuracy in several clinically relevant test cases. Compared to tracking between continuous frame for relative pose recovery, localizing every frame globally for absolute pose recovery avoids accumulated errors and the effects of failure cases, which occurs more frequently in surgical videos than ordinary tracking tasks. Low dimensional mapping in Stage I was shown to significantly improve the efficiency of image retrieval and can be used for coarse localization in challenging conditions that might be encountered in surveillance telecystoscopy. As shown in Fig. 2.29, our coarse

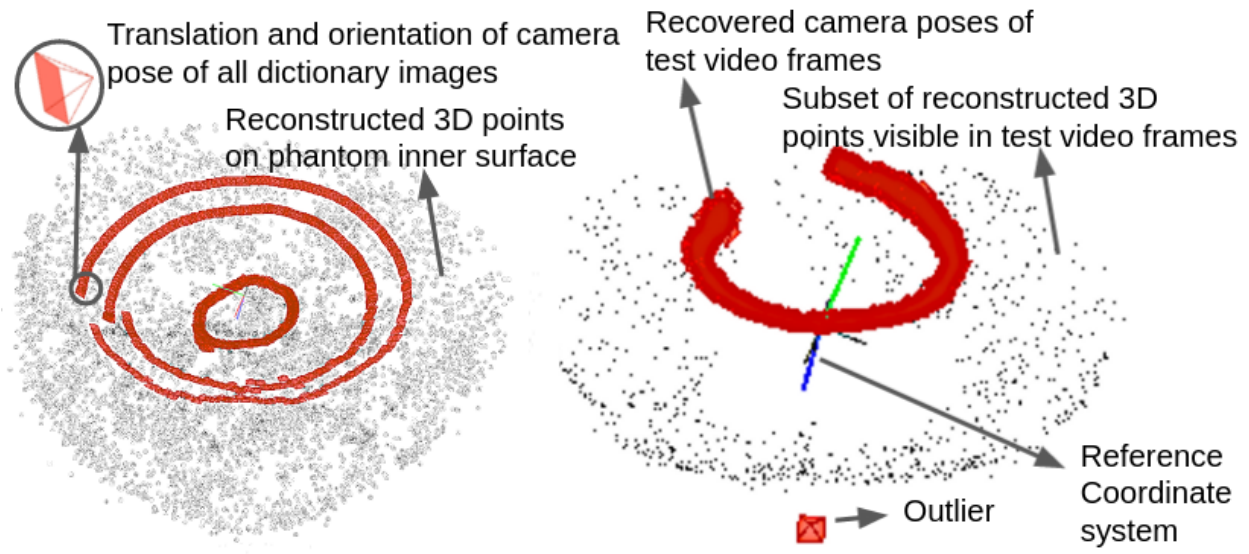


Figure 2.31: **(Left)**: Visualization of reconstructed 3D point cloud and camera poses of all dictionary images. **(Right)**: The subset of reconstructed 3D points that are visible in test video frames and the therefrom recovered camera poses of test video frames.

localization step has an mean error of less than 10° including the error in the kinematic ground truth. So the coarse localization can be independently used when the high speed is required or feature matching in Stage II fails. The coarse localization using cystoscopes with 100° FOV should provide sufficient accuracy for presenting pose estimates and maintain sufficient overlap with the prior map to teleoperators.

PCA used in our Stage I is sensitive to outliers, occlusions, and corruption in the data. Robust Principal Component Analysis (RPCA) was introduced to address this issue.[46,89] In general, RPCA is more expensive than PCA, requiring an iterative optimization to decompose the original matrix. In our pipeline, dictionary images are selected during the 3D reconstruction algorithm to be good-quality frames. With few enough outliers in the dictionary set, RPCA is not necessary. However, it is important to keep RPCA as an option for data with outliers and corruption. The distance map in Fig 2.28 shows that our dimension reduction process keeps the position relationship of the input dictionary frames in this near-clinical data while providing a more efficient searching space. If this were not the case, dark

areas would appear away from the main diagonal within one circle of scanning, indicating that the dimension reduction did not sufficiently separate disparate images within the LDS. In such a case, new images may be mis-matched to the wrong area of the bladder.

Hysteresis modeling of our cystoscope shows that image-based pose estimation is needed for providing a capable and reliable teleoperation interface for robotic cystoscopy since real time, accurate forward kinematic estimation may be difficult. To inspect the entire urothelium, urologists will deflect a flexible cystoscope against the bladder. Although this achieves viewing angles in retroflexion, this also introduces significant difficulties in estimating the pose of the scope with traditional kinematic approaches. In the 2.5D phantom case, we use cystoscope-specific kinematics to provide ground truth angulation data for the image dictionary in Fig. 2.29 after alignment, and a camera pose estimate is derived from the 3D reconstruction in the 3D phantom case. In the next step, a reliable ground truth of camera poses will be collected from extra sensors, for example, attaching electromagnetic tracking sensors on the cystoscope tip, to quantitatively evaluate the reconstructed camera pose trajectory.

We tested our approach in both 2.5D phantom and 3D phantom of the bladder. The 2.5D phantom is simple in shape so that our initial single-DOF robotic cystoscope can cover the whole phantom while recording videos that simulate cystoscopy. It is also rigid and open so that we can measure the ground truth trajectory of the camera easily and evaluate the accuracy of kinematics. The 3D phantom is an effort to evaluate our method in a environment with a more realistic shape (by using a 3D phantom made of distensible and deformable material). However, the acquisition of ground truth for camera trajectory/poses is much harder in this case because the phantom is close and we can't rely on measurement from the electromagnetic tracker due to the large error observed. Also, the manually designed vessel features on it are not realistic enough and cannot well present the robustness of our method over image degradations. Thus, we describe the experimental results on both the 2.5D and 3D phantoms to present the performance and potential of our method as clearly as possible. The scanning of 3D phantom is performed with the phantom filled with water,

which more closely simulates clinical conditions. The captured videos therefore contain bubbles and floating debris. The deformable phantom material can cause local distortions during the scanning. Perspective changes between different scans will cause the features and illuminations from the same region to appear different. With homogeneous vessel features from a larger surface in 3D phantom, there will be more local optima interference in the global localization. The spatially dense, hand-painted vessels in the 3D bladder (Fig 2.30) also provide more SIFT features than the printed human bladder image in the 2.5D phantom, thus allowing the SIFT-only method to achieve higher matching accuracy in the 3D tests than in the 2.5D tests. However, the increasing runtime factor of the SIFT-only case, from 100x to 1000x between 2.5D and 3D tests, making it hard to use in teleoperation where reasonable computational complexity is important. Although the dictionary set in the 3D case only covers a portion of the bladder phantom, an increase in the dictionary size should not greatly affect the runtime of our method.

For camera pose recovery in Stage II, we also experimented with using the 2D-2D feature correspondences between the test image and its retrieved dictionary image to calculate the transformation between the two images and then recover the camera pose of test image. We observed that using 3D-2D correspondences for camera pose recovery has better reliability than using 2D-2D correspondences. This is reasonable since the global bundle adjustment in the reconstruction step provides 3D points that are calculated to be more globally consistent with all collected images. Thus the 3D-2D correspondences are much more well-constrained and less subject to noise, compared to 2D-2D correspondences. The trajectory of the recovered test frame poses shown in Fig 2.31 qualitatively indicate the reliability of camera pose recovery from 3D-2D correspondences, as the trajectory of the source test video is a similar circle scan at a constant tip bend angle.

Future development of image-based localization using the 3D phantom can investigate new approaches to maintain robust tracking. For example, multiple dictionary images can be retrieved for each test frame and their matching relationship with the test frame can be studied to find more reliable 3D-2D correspondences. Utility of the 3D reconstruction

and real-time image matching can provide new user interfaces in teleoperation of medical robotics. Our 3D reconstruction results demonstrate reasonably accurate reconstruction of shape and texture of the bladder, which is crucial for accurate display of the bladder during teleoperation. Once camera pose of a new image is recovered, the newly acquired image can be mapped onto the 3D surface model and highlighted on the model for the operator. Not only will this help situational awareness during telecystoscopy, this could also be implemented during manual cystoscopy for training urology residents. If examined image patches are shown in contrast with unexamined areas, trainees can visualize completeness during the procedures and a real-time completeness metric can be calculated.

Additional testing is required to demonstrate efficiency and accuracy with more realistic cystoscopy videos with a wide range of bladder cancer tumors and natural anatomical variation. The experiments conducted on these phantoms provide higher image quality than a real cystoscopic video from a human bladder containing urine and water/saline. In addition, the bladder surface deformation during scanning is also not considered in the performance evaluation. When using clinical videos, the 3D reconstruction and localization performance may be affected by image degradation. With the proposed two-stage framework, both the coarse localization and camera pose recovery in our pipeline may be improved with deep-learning based approaches.[143,144] Moreover, our localization method could be especially useful when combined with other estimation technologies. For instance, if applying continuous frame tracking, our coarse localization can provide a quick and accurate estimate to regain tracking when continuous localization fails. Finally, a Kalman filter could be used to combine our global localization with continuous frame tracking and endoscope kinematics to make a more robust teleoperation system.

2.3.6 Conclusion

Our coarse localization algorithm is shown to be 100-1000x faster than a SIFT-only dictionary matching approach in the context of a two-stage camera localization pipeline that could be used for bladder cancer surveillance where 3D bladder models can be reconstructed after a

primary exam. In followup visits, our algorithm can efficiently estimate a flexible cystoscope's tip pose at around 20 Hz in bladder phantoms. We believe that our algorithm will be able to perform well in more realistic scenarios and could help make telecystoscopy a compelling option for urologists and their patients.

Chapter 3

ANALYSIS WITH DEEP LEARNING FEATURES

3.1 Eye-tracking with real-time retinal localization in AR/VR

3.1.1 Introduction

Since the eye often reaches an object of interest before the end of most head movements, integrating eye-tracker into virtual reality (VR) and augmented reality(AR) headset is highly desirable. Eye tracking could enable a new class of gaze mediated input[145] and techniques such as foveated rendering,[146] reducing the computational demands of AR/VR by providing high render quality only at the users gaze area. High accuracy eye tracking could also enables studies of how the human vestibulo-ocular system responds to virtual reality.[147] In this chapter, we present a new eye-tracker which aims to achieve the real-time tracking of the retinal movement.

Current eye tracking methods can be classified into two categories: interpolation-based and 3D model-based approaches. The interpolation-based methods assume the transformation between gaze coordinates and captured eye signals or features has a particular parametric form, such as a polynomial, or a nonparametric form like in neural networks. The 3D model-based approaches determine the geometric model of the eye, thus they can compute the gaze direction directly based on the eye features.

Interpolation-based eye tracking avoids explicitly modeling the geometry of the eye. It can be separated into invasive and non-invasive types. Invasive methods include electro-oculography (EOG) and scleral contact lens/search coil methods. Non-invasive methods are imaging-based methods. Since non-invasive methods are preferred in generalized AR/VR field, we mainly introduce the imaging based methods here. Image-based methods utilize extracted eye features such as limbus, contours, pupil, and corneal reflections to build the re-

lationship with gaze location. Most of the eye features are created under near infrared (NIR) illumination. To separate eye movements from head movements, two points of reference on the eye are needed. Since the differential movement of pupil and glints (corneal reflection) are relatively easy to find, it becomes the most popular approach for gaze estimation.[148,149] In such approaches, the detection accuracy of the pupil and glint from the eyeball surface varies with environment change, and is not consistent when the user's eyeball surface is not smooth. On the other hand, the movement trajectory of the pupil and glint is non-linear especially when the eye moves to a large angle, which increase the calibration points and the difficulty of regression. These methods currently achieve an accuracy of 0.5°-1°, while the tracking resolution of such features is around 0.7°-1°.[147,149] It is not easy to further improve the accuracy beyond the tracking resolution. Besides using features of the eye surface, retina images are also utilized for eye-tracking in medical field, such as eye-tracking scanning laser ophthalmoscopes (SLOs).[150] They leverage the scanning distortion for retinal movement estimation in small FOV high-resolution images, however this technique is designed for small saccades and SLOs are not easily integrated into a HMD.

3D model-based eye tracking uses a geometrical model of the human eye to estimate the center of the cornea, optical and visual axes of the eye and estimate the gaze coordinates as points of intersection where the visual axis meets the scene. 3D model based methods using single camera have been reported in.[151,152] Single camera systems have simple system geometry, no moving parts and fast re-acquisition capabilities. Optical and visual axes of the user's eyes are reconstructed from the centers of the pupil and glint in the captured video frames. Multi-camera methods can reach high accuracy and robustness against head movement, while they require elaborate system calibration procedures including calibration of cameras for 3D measurements, estimating positioning of LEDs, determining the geometric properties of the monitors and their relation with the cameras.[153] An accurate 3D model-based eye-tracker generally requires stereo cameras and the design is hard to integrated into VR/AR headsets.

According to the disadvantages of the image-based Interpolation methods and 3D model-

based methods, we want to exploring a new eye-tracking approaches: Utilizing the retina image directly for eye-tracking. When eyes are rotating, we can capture different regions of retina image through the pupil, then the captured frame can be localized on a large FOV retina image, as shown in Fig. 3.2 to compute the current eye sight. The full image is mosaicked by a series of captured retina images. The head movement can be compensated by the sensors in the helmet, and our research is to achieve high-accuracy gaze tracking in a certain range. This method belongs to the interpolation-based eye-tracking. The retina-based localization can reach a much higher tracking resolution compared to corneal reflection based methods and has a linear mapping relationship.

Virtual retinal display (VRD, also known as retinal scan display) for AR/VR has been proposed for a long time,[154] as shown in Fig. 3.1. To maintain the compactness of the HMD system, the retinal imaging can share most of the optic path with the retinal scan display. VRD draws a scanning display directly onto the retina, thus the scanning fiber endoscope (SFE) with a spiral scanning pattern is used to capture retina images in real time. SFE is a 2D imaging technique with the advantages of miniature probe tip and expected low cost,[155,156] and the design is shown in fig. 3.3. The SFE has a spiral scanning pattern when imaging and each pixel is imaged at different time, which can cause a large distortion on the image when the retina has a fast movement. The illumination of SFE on the target can creates reflective points on images, reducing the image quality. Meanwhile, as described in Chapter 2, the difficulty of retina localization is increased by sparse and similar retina features. Accordingly, an accurate retinal localization method to deal with scanning distortion and low quality template images is needed in the proposed eye-tracker. Different from the goal in Chapter 2, only the horizontal and vertical position in template matching is the most important information in eye-tracking, thus we can focus on the x and y translation instead of the affine transformation. It makes the problem is set as a tracking task instead of an accurate template matching or registration task. Since we want to extract deep features from the frame to make it more robust to image degradation and movement distortion, deep learning tracking methods are considered.

Visual object tracking is a basic building block in various tasks of computer vision, but many tracking methods are focusing on how to learn a robust representation of a specific target, which is not consistent with our current task. One branch of trackers is based on correlation filter, which trains a regressor by exploiting the properties of circular correlation and performing operations in the Fourier domain. It is based on finding the maximum correlation on the full image and can be leveraged in our task even our object is keep changing in each frame. In this branch, Siamese-RPN[157] achieves leading performance in the public tracking dataset. So we modified the Siamese-RPN as our retinal localization neural network. However, the performance of the network is not always reliable, having a large variance in the error distribution. On the other hand, the traditional feature point based registration approach can have a success rate around 60%. Thus we proposed a method to use Kalman filter to combine the performance of deep learning and the classic image registration method, where the result of deep learning is used to build the state transition model and image registration provides the measurement. Based on the noise distribution of the state transition model and the measurement model, we can estimate a joint probability distribution of the unknown state for each frame. The schematic of the proposed method is shown in Fig. 3.5.

Our method is validated on the synthetic data and retinal phantom movement videos imaged with the scanning fiber endoscope (SFE). The details of our dataset and its challenges are introduced in the next section. Using the retina videos, the eye tracking resolution in our system is 0.05°/pixel. Our retinal localization method currently achieves 0.68° mean error not considering the annotation variation. Compared to the classic pupil-glint methods which have a low tracking resolution, we hypothesize that retinal-based eye tracking accuracy will greatly improve in future development.

3.1.2 Data Characteristics

As described above, SFE has a spiral scanning pattern from center to periphery and the full frame is imaged ring by ring. When the target is moving, the rings scanned at different time

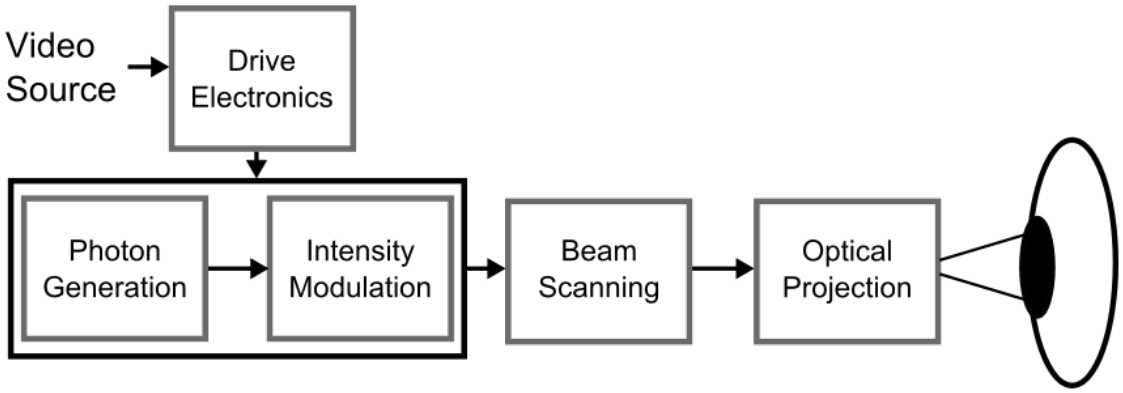


Figure 3.1: Optic path of the virtual retinal display.

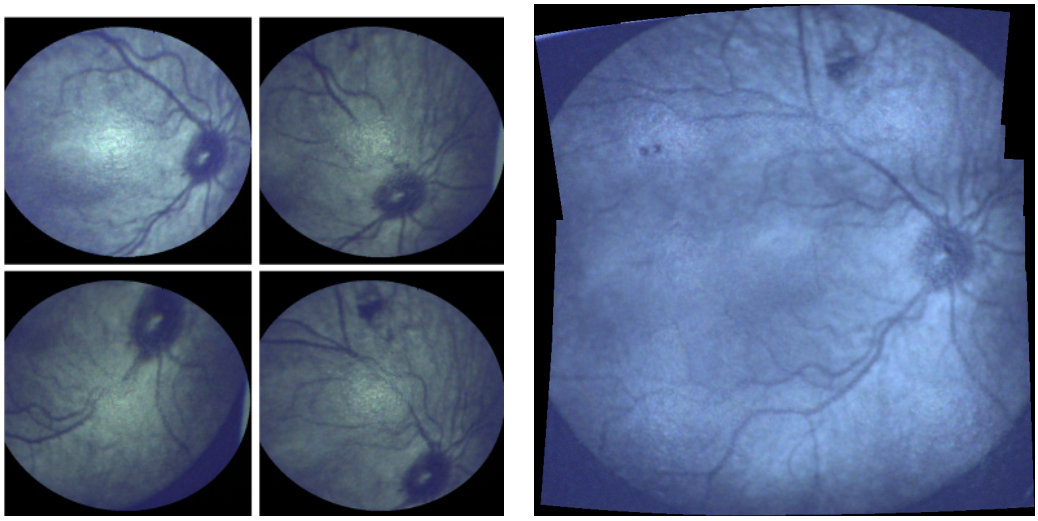


Figure 3.2: Example retina frames of the SFE and the mosaicked large FOV baseline image.

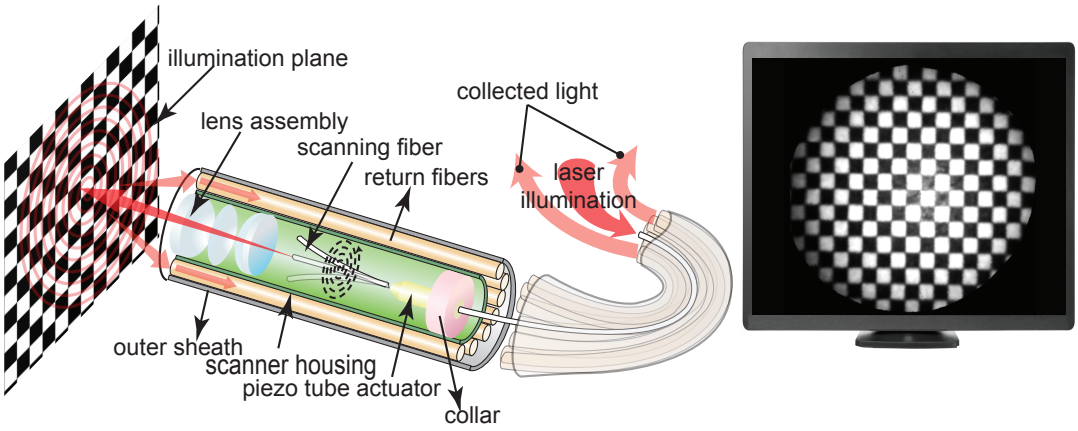


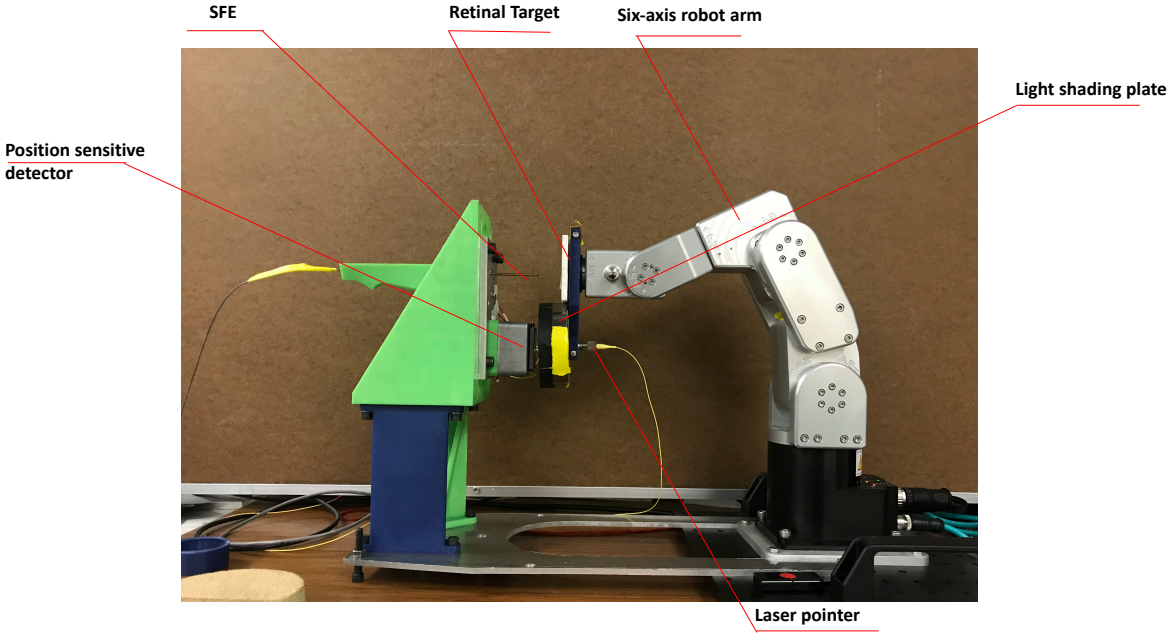
Figure 3.3: Design of scanning laser endoscope.

are from different regions, which creates movement distortions in the video frame. We take the imaging position on the retina as the ground truth when the frame is completed, thus the outer rings in each frame are closer to the ground truth.

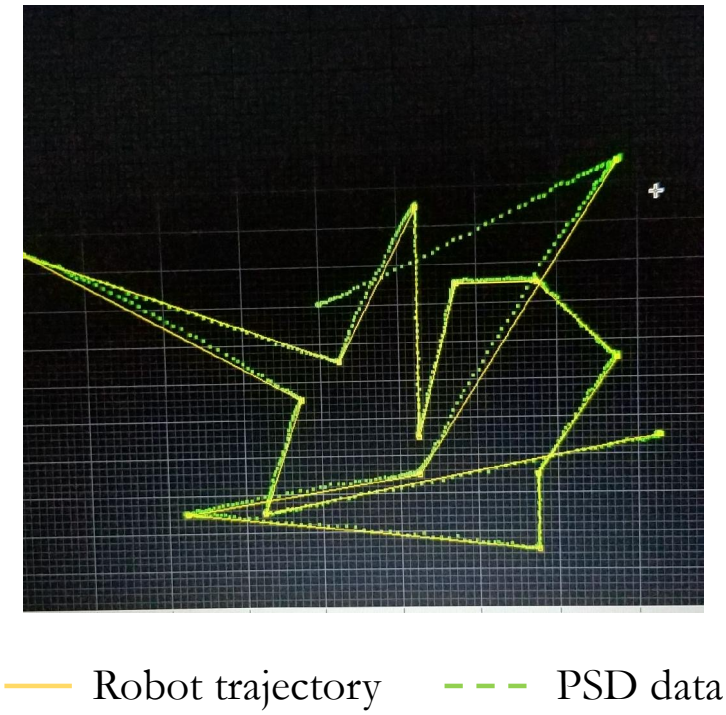
Fig. 3.2 (a) shows example of captured retinal frames, and Fig. 3.2 (b) is the image mosaicked from a series of frames. We can see from the donut-shaped optic disc that the images have movement distortion. Note that the retina image has many regions with similar background, the imaging has low quality with local distortions on the still frame, which increase the difficulty of localization.

3.1.3 Proposed Approach

Given a large FOV mosaicked image as reference, our task is the real-time localization of the captured SFE frames onto the search image as shown in Fig. 3.2 (Right). Because of the challenges of data, we use deep learning method to extract representative deep features for analysis. However, the neural network has uncertainty and deep features are not always reliable, thus image registration method is used to compensate the performance of deep learning. On the other hand, the result of image registration is also noisy because of the data challenges. As described above, two process are combined with the Kalman filter, where



(a)



(b)

Figure 3.4: (a) The robotic platform for labelled data collection; (b) An example of the trajectories the robot followed and the record of PSD.

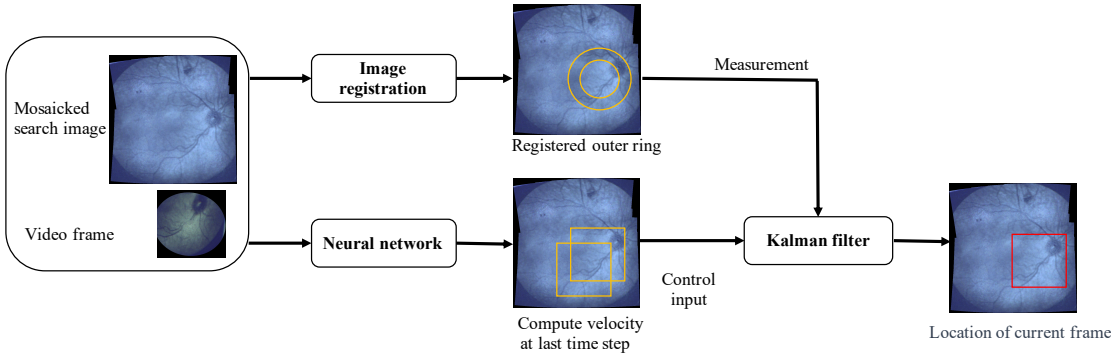


Figure 3.5: Schematic of proposed real-time localization method. Kalman filter is used to combine the performance of deep learning and the classic image registration method.

the deep learning results are embedded in the transition model and registration results are taken as the measurement in Kalman filter. The Kalman filter requirements of linear Markov model and additive Gaussian noise are satisfied in our case. In this section, we introduce the form of our state transition model and measurement in the Kalman filter respectively.

State Transition Model with Deep Learning

Kalman filtering, also known as linear quadratic estimation (LQE), is an algorithm that uses a series of measurements observed over time, containing statistical noise and other inaccuracies, and produces estimates of unknown variables that tend to be more accurate than those based on a single measurement alone, by estimating a joint probability distribution over the variables for each time frame.[158] Kalman filters are based on linear dynamical systems discretized in the time domain. They are modeled on a Markov chain built on linear operators perturbed by errors that may include Gaussian noise. The state of the system is represented as a vector of real numbers. The Kalman filter model assumes the true state at time k is evolved from the state at $(k-1)$ according to:

$$x_k = F_k x_{k-1} + B_k u_k + w_k, \tag{3.1}$$

where F_k is the state transition model which is applied to the previous state $x_k - 1$; B_k is the control-input model which is applied to the control vector u_k; w_k is the process noise which is assumed to be drawn from a zero mean multivariate normal distribution with covariance Q_k: $w_k \sim N(0, Q_k)$.

After the state transition model, the measurement is obtained at the current time:

$$z_k = H_k x_k + v_k, \tag{3.2}$$

where H_k maps the state to the measurement and v_k is the measurement noise. Similar with the transition model, $w_k \sim N(0, R_k)$.

In the proposed method, the transition model is formed as follows:

$$\begin{bmatrix} X_k \\ Y_k \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} * \begin{bmatrix} X_{k-1} \\ Y_{k-1} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} * \begin{bmatrix} \dot{X_{k-1}} \\ \dot{Y_{k-1}} \end{bmatrix} + w_k, \tag{3.3}$$

X_k, Y_k represents the position state at time k in x and y directions. w_k is the process noise drawn from a zero mean multivariate normal distribution. $\dot{x_{k-1}}$, $\dot{y_{k-1}}$ forms the control vector of the first order state estimation model. It is the velocity within a time unit computed from the difference between the deep neural network results at time k and $k-1$. The proposed formation allows us to embed the deep learning into a classic Kalman filter model. Here one time step is the duration between continuous frames.

The deep learning framework we use is modified from the Siamese RPN.[157] The network structure is shown in Fig. 3.7. The network consists of a Siamese subnetwork for feature extraction of the full image and the current frame, and a region proposal (RP) subnetwork to generate proposal of the frame regions on the search image feature map. The RP subnetwork is inspired by RPN in object detection.[159] The basic idea of RPN is setting uniform anchors on the input image, and for each anchor, several different scale and ratio anchor boxes are set, as shown in Fig. 3.6. Then the network will predict several candidate anchor boxes as the potential object (proposal) in the output. The region proposal subnetwork consists

of a pair-wise correlation section and a supervision section. The supervision section has classification branch and proposal regression branch. If there are k anchors, network needs to output $2k$ channels for classification, indicating the positive and negative score for each anchor. It needs $4k$ channels for regression, indicating the $x, y, width, height$ adjustment for the candidate anchor boxes compared to the ground truth. So the pair-wise correlation section first increase the channels of the template feature to $2k$ and $4k$ times for two branches. The full image feature is also split into two branches by two convolution layers but keeping the channels unchanged. Then the correlation is computed on both the classification branch and the regression branch.

Specifically, in our implementation, Alexnet[160] is used to extract the deep features of the frame and search image. The structure of Alexnet is shown in Fig. 3.8. After feature extraction, the frame feature map is converted into two feature maps with convolution layers for different branches, then two corresponding response maps are created by the convolution of each frame feature with the search image feature (correlation layers). One response map is used for the target region/non-target (positive/negative) region classification, another response map predicts the position refinement at each positive position.

In the training process, the loss function is the same as in Faster R-CNN:[159]

$$loss = L_{cls} + \lambda L_{reg}, \tag{3.4}$$

where λ is a hyper-parameter to balance the two parts. Classification loss is the cross-entropy loss and smooth L1 loss with normalized coordinates is used in regression branch. Let A_x, A_y, A_w, A_h denote center point and shape of the anchor boxes and let T_x, T_y, T_w, T_h denote those of the ground truth boxes, the normalized distance is:

$$\delta[0] = \frac{T_x - A_x}{A_w}, \delta[1] = \frac{T_y - A_y}{A_h}, \delta[2] = ln\frac{T_w}{A_w}, \delta[3] = ln\frac{T_h}{A_h} \tag{3.5}$$

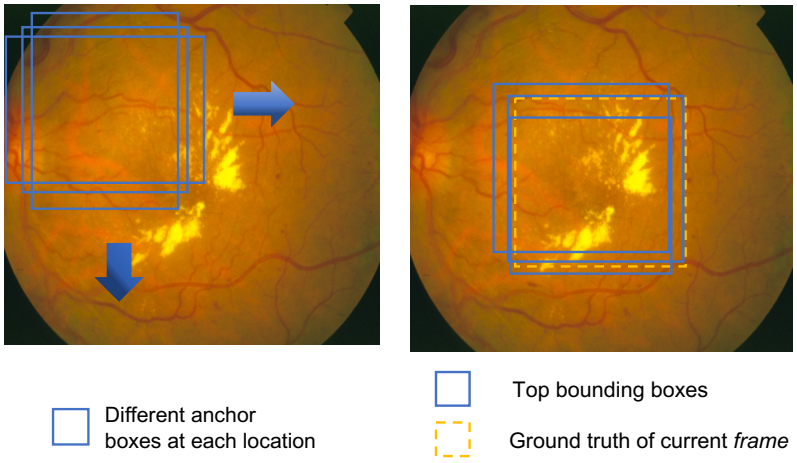


Figure 3.6: Anchor setting of RPN. (a) shows at each anchor point, k different anchors with different scale / ratio are set. (b) indicates the network will rank all the anchors and provide the top several ones with the highest probability, and the difference in location and size between the selected anchor boxes and the ground truth.

Then they pass through smooth L1 loss:

$$smooth_{L_1}(x,\sigma) = \begin{cases} = & 0.5\sigma^2 x^2, |x| < \dfrac{1}{\sigma^2} \\ = & |x| - \dfrac{1}{2\sigma^2}, otherwise \end{cases} \tag{3.6}$$

and L_{reg} is:

$$L_{reg} = \sum_{i=0}^{3} smooth_{L_1}(\delta[i],\sigma). \tag{3.7}$$

Different from learning robust representations of a specific object in Siamese RPN, we localized different templates on the same search image. The deep feature of the search image is saved and repeatedly used after the training process. Since the imaging scale will not change much in HMD, we focus on the target position in x and y instead of the bounding box with adjustable height and width. A small weight is assigned to the height and width related loss considering the influence of the motion distortion can enlarge the region of the

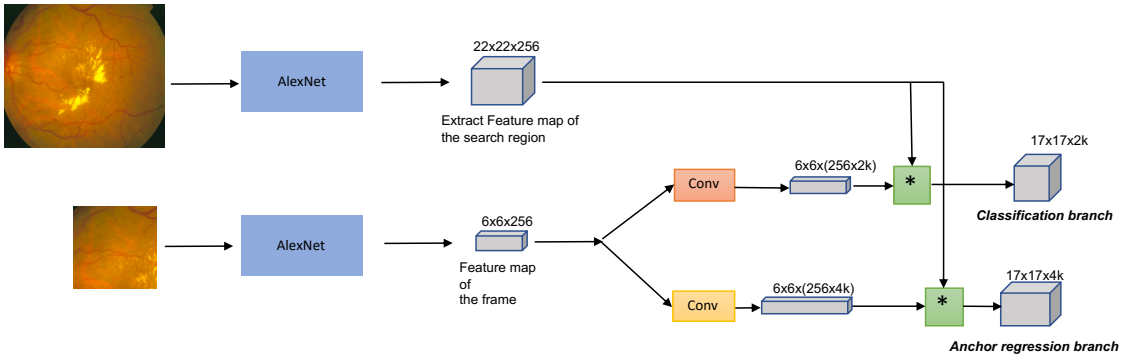


Figure 3.7: Structure of the deep neural network for retinal localization.

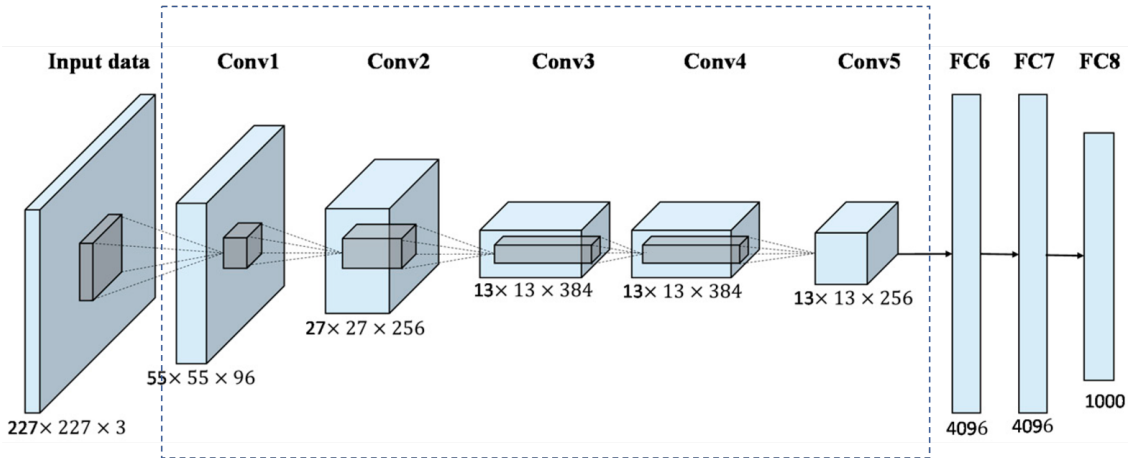


Figure 3.8: Structure of the Alexnet. The convolution layers in the dash box is used for deep feature extraction.

frame on the search image. After training, the deep feature map of every full image can be saved to avoid repeated computation, which can make the localization more efficient, especially in a mobile device in the future.

Measurement with Outer Ring Registration

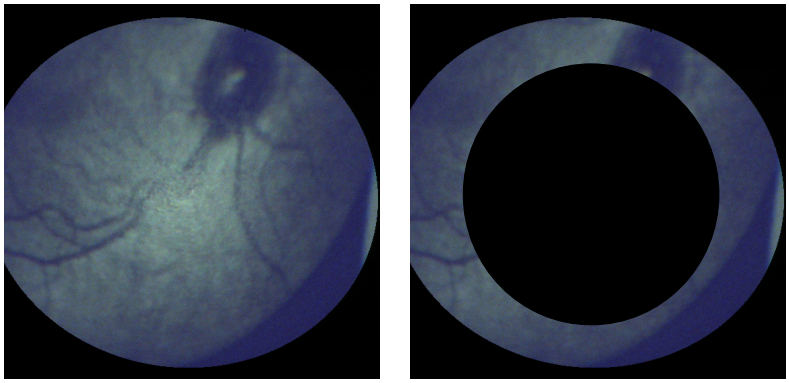
In Kalman filter, the measurement is obtained at the current time:

$$z_k = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} * \begin{bmatrix} X_k \\ Y_k \end{bmatrix} + v_k, \tag{3.8}$$

where z_k is the measurement obtained by the image registration, and v_k is the measurement noise similar to w_k.

The measurement of the Kalman filter is obtained by the separate image registration process. Considering the accuracy and speed, SIFT is selected as the image registration method. As described before, the SFE has a spiral scanning pattern from the inside out, thus there is a start and end time point for imaging the entire frame. When the target is moving, the PSD ground truth is corresponding to the target location at the end time point, which is indicated by the outer ring of the captured frame (Fig. 3.9). If the whole frame is used to do the registration with SIFT, the distortion caused by the mismatch between inner rings and outer ring will be counted, which has a different mapped center point position with the ground truth. Since we only care about the outer ring's position in a frame, only the outer ring is registered onto the full image. However, the outer ring only provides little feature information, so it is hard to match it globally. To reduce the interference of similar background and few features of outer ring registration, the image registration includes two steps: coarse registration of the whole frame and fine registration with the outer ring only.

In the coarse localization, we detect feature points from two images and register the frame f to the corresponding regions $\tilde{f}$ on the search image. In the outer ring registration, the feature points within the enlarged region around $\tilde{f}$ on the search image are selected, and they are rematched with the feature points falling into the outer ring region on the frame. A new transformation matrix that computed by registering these two set of feature points again is the registration result for outer ring. Using an enlarged region improves the robustness of the algorithm when the matched featured points in coarse registration are concentrated in



(a) (b)

Figure 3.9: (a) Original frame. (b) Outer ring.

the inside area. This method also avoids repeated computation of feature points.

In both of the coarse and fine registration, the success registration can be indicated by the number of matched feature points. When the outer ring registration fails because of insufficient feature points, we use the registration result of the whole frame as the measurement at current time. If the whole frame registration is not successful, the retinal localization result depends only on the state transition model and no observation update.

3.1.4 Experiments

Experiments are performed on two datasets: the synthetic retina movement videos and SFE videos introduced above. We compares the performance of the proposed tracking method with Kalman filter with using the deep learning only.

Synthetic retina movement data

The synthetic data is generated from the public retina dataset STARE,[40] which has been introduced in Chapter 2. Three hundred of retina images are selected as the full image, and several retina movement videos are generated for each full image. The number overall movement frames is 36000, including data augmentation. We add different degradation on

Table 3.1: Degradation levels for test dataset.

Distortion level	0	1	2	3	4
Gaussian var.	-	0.002	0.003	0.004	0.006
Rotation	-	4°	6°	8°	12°
Shear	-	2°	3°	4°	6°
Rescale	-	0.8	0.9	1.1	1.2

the generated video frames: 1) Gaussian noise with $mean = 0$ and variance randomly selected from $[0.001, 0.005]$; 2) Rotation angle ranges from $-10°$ to $10°$; 3) Shear angle ranges from $-5°$ to $5°$; 4) Scale change from 0.8 to 1.2. The test data includes 250 video frames. We set four different degradation levels from low to high for test dataset and to see the performance of the trained model on different test set. As shown in Table 3.1, including the clean data (level 0), there are five test data groups. Fig. 3.11 shows the error distribution in different degradation levels with our Kalman filter based method and using the deep learning part only. We can see the mean and standard deviation of each case are reduced with Kalman filter compared against using the deep learning part only. Combining the outer ring based registration helps us to obtain a more reliable tracking performance. Using Kalman filter has an accuracy 0.63° under the largest degradation, which is close to the upper limit of current pupil-glint eye tracking methods with an accuracy range of $0.5° - 1°$.

Retina phantom videos imaged by SFE

As described before, we use SFE to scan the retina target. For supervised training of the localization neural network, the collected video frames should have corresponding ground truth: the x and y coordinate position on the retina. To achieve the automatic annotation, we build a data collection platform.

Data Collection Platform

The data collection platform includes SFE, a robot arm for simulating retina movement and a position sensitive detector (PSD) for real-time position recording. The robot arm is

Meca500 from Mecademic Inc. Meca500 is a six-axis industrial robot arm that is relatively easy to use, robust and lightweight, The robot is, however, a precision device with rapidly moving parts. The amplitude of eye saccade is the angular distance the eye travels during the movement. For amplitudes up to 15 or 20°, the velocity of a saccade linearly depends on the amplitude. Head-fixed saccades can have amplitudes of up to 90°, but in normal conditions saccades are far smaller, and any shift of gaze larger than about 20° is accompanied by a head movement. When the amplitude is 20°, the speed reaches 500°/s. Converting to the target movement speed is around 300mm/s. The highest speed of the robot arm can reach 500mm/s. It allows us to use the robot arm to simulated the fast eye movement.

In the platform design, the SFE is fixed in front of the robot arm tip, where the retina target is attached. A fiber laser is fixed under the target for PSD input, which will repeat the robot movement trajectory for PSD recording. The steps of the data collection are as follows: 1) Start to record the SFE imaging video. 2) Start to record the PSD data. 3) Flash the LED both on retina target and PSD as start signal. 3) The robot arm operates the retina movement trajectory. 4) Flash the LED both on retina target and PSD as the end signal. 5) Stop data recording. The start and end LED flash is used for the SFE and PSD data synchronization. The imaging frequency of SFE is 29.7183Hz. We use a high speed video card to collect the SFE images to avoid dropping frames, and its frequency is around 72Hz. The frequency of PSD data collection is 4000Hz.

Data Pre-processing

The data we collect from the platform includes the high-frequency SFE frames and the corresponding PSD position data. Since the SFE frames frequency speed is much higher than the real imaging speed, there will be many repeated SFE frames. The first step of data pre-processing is to delete those repeated frames by setting a threshold (based on summary of grayscales) to recognize the repeated continuous images. After that, the frequency of SFE frames will be reduced to 30Hz. Secondly, the start and end LED flash frames are detected, with the corresponding LED flash PSD points, which have a large power than normal PSD data. We assume the data collection time of SFE and PSD is uniform, so the time for each

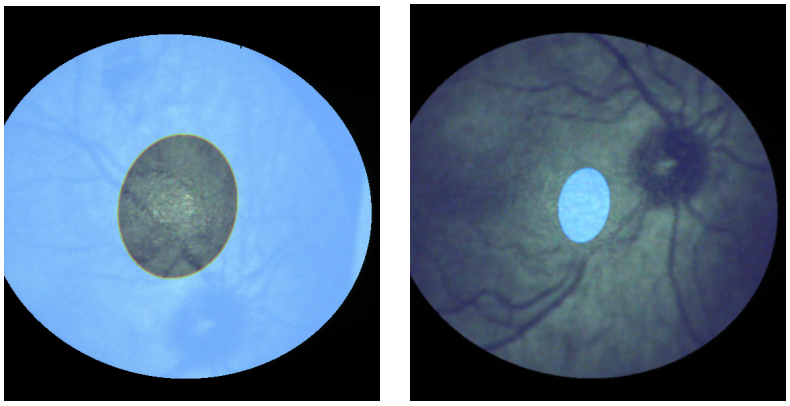


Figure 3.10: Examples of LED flash on the SFE frame.

SFE image can be matched with the PSD data within the time between two LED flashes. Now each frame has its ground truth in millimeter, and we need to transfer the ground truth into pixels on the full image. We sample 20 frames have minimal distortions and manually register them with the full image. The we can get a series of data pair between the PSD ground truth in millimeter and the center point coordinate on the full image in pixels. A linear model is fitted with these data points to compute the PSD ground truth in pixel for all frames.

The annotation in the current setup has a mean error of 0.35°. The error is contributed by the robot precision, PSD precision, LED flash (The flash will stop before or after one entire frame has been scanned as shown in Fig. 3.10. It will lead to the time mismatch within one frame if we take it as the start and end signal), and the model fitting error.

Results

For the SFE data, we collect movement videos for one retina at the current stage, and the number of frames are around 7000. 400 frames are used for test and the rest are training dataset. The Cumulative distribution function (CDF) of the test errors is shown in Fig. 3.12. We can see using Kalman filter reduces the mean and variance of the test error distribution, and there are outliers over 5° in using deep learning only. Excluding the influence of the annotation error, the mean error of our method is 0.68°. The speed can reach 72fps with the

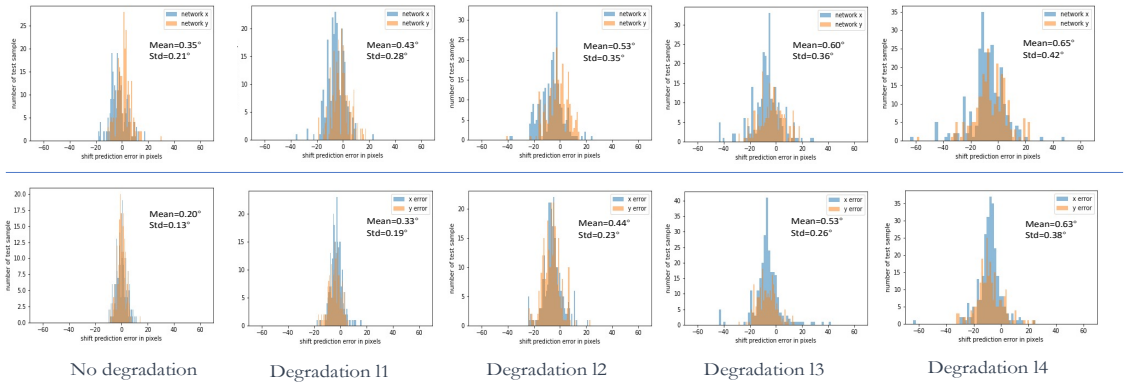


Figure 3.11: Test error distribution over the simulated data with different degradations. The first row is the performance of the neural network and the second row is the proposed method's performance with Kalman filter. The marked mean and std is converted to degrees.

GPU of Titan RTX.

3.1.5 Conclusion and Future Work

We present the application of retina-based eye tracking for HMD and a novel real-time localization method using Kalman filter to combine the performance of deep learning and image registration. To the best of our knowledge, this is the first systematic discussion of embedding retina tracking in AR/VR headset and providing algorithm solutions.

We expect to further improve the accuracy and generalize the model trained on SFE videos by enlarging the dataset from different user's retina, then evaluate our method on in vivo cases in the future. To improve the measurement in Kalman filter, deep learning can also be used to learn more robust feature points in the image registration part.[161] We predict the accuracy of the widely used pupil-glint methods may have reached its tracking resolution limit using current sensors, whereas using retina videos has the capacity for improving their degree of accuracy. Retina based eye tracking also provides a more precise gaze estimation model by computing the fovea position, and will enable more applications in the medical field such as the retinal diagnosis or surgery guidance.[1]

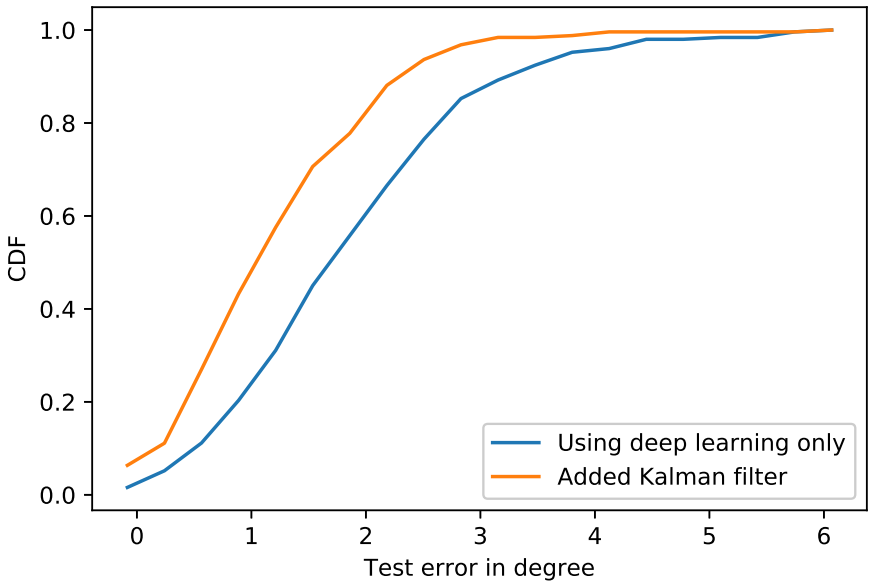


Figure 3.12: CDF of the retina tracking errors in degrees over 400 frames, and the annotation error is included in the CDF.

3.2 Synergistic network learning and label correction for noise-robust training

3.2.1 Introduction

Deep learning has shown very impressive performance on various vision problems. However, a practical challenge for deep learning state-of-the-art models is that they rely on large amounts of clean, annotated data.[160,162–164] Collecting such data sets is expensive or time-consuming. Large training datasets almost always contain examples with inaccurate or incorrect labels, resulting in overfitting to noisy samples and poorer model performance.[165–168] There are many classic methods to prevent noise overfitting such as dropout[169] and early stop,[170] which are heuristic for noisy label learning.

A large number of algorithms have been developed for learning with noisy labels. Small loss selection recently achieved great success on noise-robust deep learning following the widely used criterion: DNNs tend to learn simple patterns first, then gradually memorize all samples.[171–173] These methods treat samples with small training loss as clean ones. During the training process, clean or confident instances are selected to update the model parameters. For instance, MentorNet[174] trains a Teacher-Net to provide curriculum, in the form of weights on the training samples, to select clean samples to guide the training of the Student-Net. Co-teaching[175] uses two networks to determine clean samples in their mini-batches separately and exchange the update information with the other network. Inspired from Co-teaching, JoCoR[176] also uses two networks, while they combine the "agreement strategy" from semi-supervised learning into noise label tasks, which uses a joint loss to make their predictions agree. The instances with small joint loss are selected for the back-propagation. Experiments in JoCoR showed that it is more effective than co-teaching. However, not all clean examples can be selected by the network with the small loss strategy. When the noise rate is high, the selection would further decrease the number of effective training samples in each batch. Currently, there are relabelling methods learning network parameters and inferring the ground-truth labels simultaneously without any clean dataset, such as joint optimization[177] and PENCIL.[178] They use all data for learning so noisy examples can be an

interference in this process.

In this work, motivated by increasing the effective training samples, we proposed a framework that combines the noise correction with small loss selection methods to update the network parameters and noisy labels iteratively (Fig. 3.13). Each time we train relatively robust models with small loss examples, and correct the noisy labels which the current networks are confident on. In the training process, we are inspired by JoCoR to use "agreement strategy" to train two networks with a joint loss. Our contributions are as follows.

(a) We proposed a robust learning framework for both network parameters and ground truth labels to handle noise label tasks. Our method is independent of the backbone network structure and does not need an auxiliary clean dataset. Training the network with small loss selection and updating the label of confident examples make the iterative learning process robust. To the best of our knowledge, it is the first method in this line. (b) We prove the effectiveness of training a stable model over current dataset before each label correction step. It performs better than conducting weights and label update at each iteration. (c) We conduct extensive experiments on both synthetic and real-world noisy datasets and our method achieves state-of-the-art accuracy.

3.2.2 Related Work on label noise

Small-loss selection is one of the training strategy in data-reweighting. These studies select samples that are expected to be clean or remove noisy ones.[179] Based on the small-loss criterion described above, the model gradually converges to a good classifier such that the mis-labeled training samples exhibit unusually high loss values during training.[180] The underlying idea of small-loss selection is to select or give larger weights to these easier samples, hence they more likely to have correct labels. For instance, MentorNet[174] trains a Teacher-Net to provide curriculum, in the form of weights on the training samples, to select clean samples to guide the training of the Student-Net. In Co-teaching,[175] they used two networks to determine clean samples in their mini-batches seperately and exchange the update information with the other network. JoCoR[176] also use two networks, while they proposed one

joint loss for two networks. The instances with small joint loss then are selected to do the back-propagation. Experiments in JoCoR showed that it is more effective than co-teaching. Although selecting clean examples can avoid the interference of noisy labels, many difficult samples will be deleted which are important to algorithm's accuracy.[181] On the other hand, simply dropping the noise-labelled examples prevents the network from maximizing the use of the existing data.

Relabelling is used to reassign the labels of noisy instances. Relabeling has two settings: including a small clean dataset and no such a dataset. In the first category, a multi-task network is proposed: one model is trained on clean sets to correct the noisy data, and another model is trained on a merged clean & relabeled dataset.[182] In the second category, the noisy labels are updated based on the prediction of CNNs,[177,183,184] and an end-to-end framework is proposed to update both network parameters and label estimations as label distributions.[178] This category has more applications in the real world since we cannot observe a clean set in most cases. Our method is also motivated by the network-based relabelling.

Noise rate estimation introduces the noise transition matrix T to correct the predictions.[185,186] Goldberger et al.[187] added an extra layer at the end of a backbone CNN that simulated the noise transition matrix. F-correction[186] used a two-step solution to heuristically estimate the noise transition matrix. In these approaches, the quality of noise rate estimation is a critical factor for improving robustness. However, noise rate estimation is challenging, especially on datasets with a large number of classes.[176]

Robust loss function is widely investigated to deal with noisy labels. For example, Ghosh et al.[188] proves that mean absolute error (MAE) is a noise-robust loss for CNNs. Accordingly, GCE[189] combines MAE and the cross entropy loss to generalize a better loss function for noise handling. Reed et al.[190] introduces 'soft' and 'hard' loss functions based on bootstrapping. F-correct[186] proposes a risk minimization method to learn neural networks by estimating label corruption probabilities. Robust loss functions achieve some success, however, they cannot perform well on challenging noisy datasets.

3.2.3 Proposed Approach

Notation

For multi-class classification with C classes, we have a dataset D with N samples. $D = \{x_i, y_i\}_{i=1}^{N}$, where x_i is the ith example with its label as $y_i \in [1, ..., C]$. The set of examples and labels are denoted as X and Y. The neural network is denoted by $f(x, \theta)$, with the output of the final layer (C-class softmax layer): $\mathbf{p} = [p_1, ..., p_C]$. $\mathbf{p}$ is the predicted probability for each example x.

During network training in a standard classification task, a loss function $\mathcal{L}$ is used to measure the distance between $\mathbf{p}$ and the label y. The network parameters θ are learned by optimizing $\mathcal{L}$ by gradient descent methods. In our study, the given label Y is noisy, and our task is to jointly optimize θ and labels Y. The learned label set is denoted as $\hat{Y}$.

Joint Training with Small Loss Selection

According to the agreement maximization principles,[176,191,192] different models would agree on labels of most examples and are unlikely to agree on incorrect labels. In the process of updating θ, two different classifiers are encouraged to make predictions closer to each other with a regularization term to reduce divergence. Specifically, joint training uses a joint loss to train two networks (same structure with different initialization) simultaneously. The loss function is composed of supervised learning loss and agreement loss as shown in Eq. 3.9,

$$\mathcal{L}_{joint} = (1 - \lambda) * \mathcal{L}_{sup} + \lambda * \mathcal{L}_{agr}, \tag{3.9}$$

where λ is a hyperparameter for linear combination and it decreases with the label noise rate. When the label noise is high, we select a high λ to trust the "agreement" between two networks instead of the supervised learning loss. With our label update process providing cleaner labels, λ will be gradually reduced and we rely more on supervised learning.

The supervised learning loss $\mathcal{L}_{sup}$ is the sum of the classification loss of two networks,

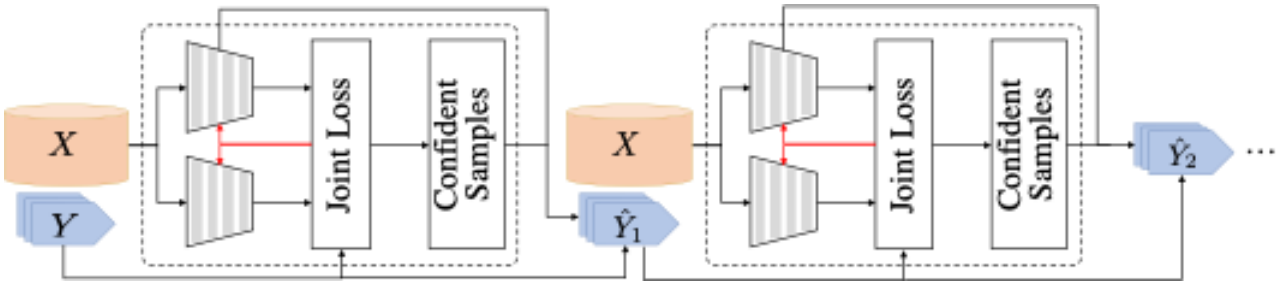


Figure 3.13: The iterative network learning and label correction of proposed method. Two models are trained with a joint loss and confident samples are selected for relabelling.

which is the Cross-Entropy between predictions and labels. p_1 and p_2 are the prediction outputs of two networks, respectively.

$$\begin{aligned}\mathcal{L}_{sup}(x_i, y_i) &= \mathcal{L}_{C1}(x_i, y_i) + \mathcal{L}_{C2}(x_i, y_i) \\ &= -\sum_{i=1}^{N}\sum_{c=1}^{C} y_i log(p_1^c(x_i) - \\ &\sum_{i=1}^{N}\sum_{c=1}^{C} y_i log(p_2^c(x_i)).\end{aligned} \tag{3.10}$$

The agreement loss is the term to reduce divergence between two classifiers which is the Jensen-Shannon divergence in Eq. 3.12, where $\mathcal{D}_{KL}(p_1||p_2)$ is the Kullback–Leibler divergence.

$$\mathcal{L}_{agr} = \mathcal{D}_{KL}(p_1||p_2) + \mathcal{D}_{KL}(p_2||p_1), \tag{3.11}$$

where

$$\mathcal{D}_{KL}(p_1||p_2) = \sum_{i=1}^{N}\sum_{c=1}^{C} p_1^c(x_i) log\frac{p_1^c(x_i)}{p_2^c(x_i)}. \tag{3.12}$$

Small loss examples are selected to do the back-propagation in each iteration and a selection rate $R(t)$ is used to control the portion. $R(t)$ depends on the iteration t and its schedule is introduced in Section 2.4.

Learning with label correction

From the view of agreement maximization principles that different models are unlikely to agree on incorrect labels, the joint loss would be high with both high supervised learning loss and agreement loss over noisy labelled examples. Given this assumption, we obtain clean labels by updating labels in the direction to decrease Eq. 3.13.

$$\min_{\theta_1,\theta_2,Y} \mathcal{L}_{joint}(\theta_1, \theta_2, Y|X). \tag{3.13}$$

In the proposed learning framework, network parameters θ_1, θ_2 and class labels Y are alternatively updated: 1) Updating θ with fixed Y: With the joint loss in Eq. 3.9, we update θ_1 and θ_2 with small loss selection. 2) Updating Y with fixed θ_1, θ_2: After relatively robust models are trained on the current dataset, part of the noisy labels are corrected.

When updating the labels Y, we want to select the labels which are noisy and the networks are confident to update. The examples with small agreement loss are considered as confident examples so we select the subset $D_{correction}$ with the least agreement loss based on the label correction rate:

$$D_{confident} = argmin_{D':|D'|\geq C(k)|D|}\mathcal{L}_{agr}(D'), \tag{3.14}$$

where $C(k)$ is the label correction rate which changes with the k^{th} label correction. Among the confident examples, we take the noisy ones whose predicted class distributions have large difference with the current labels. Similarly, this subset D_{noisy} can be selected with the large supervised learning loss according to the label correction rate $C(k)$:

$$D_{noisy} = argmax_{D':|D'|\leq C(k)|D|}\mathcal{L}_{sup}(D'). \tag{3.15}$$

The subset for label correction $D_{correction}$ will be the intersection of D_{noisy} and $D_{confident}$:

$$D_{correction} = D_{confident} \cap D_{noisy}. \tag{3.16}$$

Within the selected examples, if two networks have the same predictions on one example, the current label is updated with the predicted class, otherwise it remains unchanged.

Instead of updating labels and parameters simultaneously in each iteration, we train two robust models on the current dataset before every label update step to make the correction more reliable. Since labels are not continuously updated, there can be a large amount of changes on the dataset after correction. We set a threshold $C_{restart}$ for label correction rate $C(k)$ as a trigger for restarting training. When $C(k) > C_{restart}$, a large portion of labels will be changed and we retrain two networks from scratch with the new labelled dataset.

We can use any deep neural network as the backbone network, and then equip it with the joint training and label correction to handle learning problems with noisy labels. After the networks have been fully trained, the label correction parts are not needed. The backbone networks alone can perform prediction for test examples and indicates ambiguous ones when two networks disagree.

Algorithm 4 presents the steps of proposed framework.

Training implementation

The label correction rate $C(k) = \tau/2 * 1/k$ since as the number of correction time k increases, fewer examples are noise-labelled. The raw noise ratio τ decreases after label update according to the size of last $D_{correction}$:

$$\tau_k = \tau_{k-1} - \frac{|D_{correction}|}{|D|}. \tag{3.17}$$

The schedule of $R(t) = 1 - min\left\{\frac{t}{T_k}\tau, \tau\right\}$, where t is the iteration and $T_k = 10$ for CIFAR-10 and CIFAR-100, $T_k = 5$ for Clothing1M.

The training of our method is implemented through two stages: (1) Iterative optimization of network parameters θ_1, θ_2 and labels $\hat{Y}$. (2) Fine-tuning θ_1 and θ_2 with fixed labels $\hat{Y}$.

Algorithm 4: Learning with Label Correction

Network f with $\Theta = \Theta_1, \Theta_2$, learning rate η, noise rate τ, epoch T_k and T_{max}, label update epochs T_{update}, iteration I_{max}, label correction threshold for restart training $C_{restart}$ Θ_1, Θ_2

for $t = 1, ...T_{max}$ **do**

  $R(t) = 1 - min\left\{\frac{t}{T_k}\tau, \tau\right\}$;

  for *n = 1,...,*I_{max} **do**

    Fetch mini-batch D_n from D;

    $p_1 = f(x, \Theta_1), \forall x \in D_n$;

    $p_2 = f(x, \Theta_2), \forall x \in D_n$;

    Calculate joint loss l_{joint};

    Obtain small loss sets $\hat{D_n}$ with $R(t)$ and corresponding loss L;

    Update Θ: $\Theta = \Theta - \eta \nabla L$;

  end

  if $t = T_{update}(k)$ **then**

    $C(k) = \tau/2 * 1/k$;

    Obtain $D_{correction}$ with $C(k)$;

    Update labels of $D_{correction}$ with p_1, p_2;

    $\tau = \tau - \frac{|D_{correction}|}{|D|}$;

    if $C(k) > C_{restart}$ **then**

      Restart training with the new dataset;

    end

  end

end

3.2.4 Experiments

In this section we first compare our method with some state-of-the-art approaches, then analyze the impact of label update intervals and retraining by ablation study. The experiments are conducted on both synthetic and real-world datasets and achieves state-of-art performance.

Datasets

Our method is demonstrated on MNIST, CIFAR-10, CIFAR-100, and Clothing1M,[193] using the PyTorch framework. We corrupted synthetic datasets with symmetric and asymmetric noise manually. The asymmetric noise is to simulate that labellers may make mistakes only within very similar classes.

MNIST: Following the work of reed2014training and patrini2017making, we use a label transition matrix Q to flip clean label y to noise label $\hat{y}$. In the symmetric noise setup, label noise is uniformly distributed among all categories, and the label noise ratio is $\tau \in [0, 1]$. For each example, the noise-contaminated label has $1 - \tau$ probability to remain correct, but has τ probability to be flipped uniformly with other $c - 1$ labels.

As for asymmetric noise, similar to JoCoR the noisy labels were generated by mapping $7 \rightarrow 1$, $2 \rightarrow 7$, $3 \rightarrow 8$ and $5 \leftrightarrow 6$ with probability τ.

CIFAR-10: The symmetric noise setting is the same as that in MNIST. Note that in the symmetric noise of PENCIL on all dataset, the label of one example has $1 - \tau$ probability to remain correct while has τ probability to be drawn uniformly from the c labels. The ratio of overall correct labels is larger than τ in this case. Besides using different network backbone, this is another main reason that our implementation of PENCIL performs worse than it reported. As for asymmetric noise, the noisy labels were generated by mapping $truck \rightarrow automobile$, $deer \rightarrow horse$, $bird \rightarrow airplane$ and $cat \leftrightarrow dog$[178] with probability τ.

CIFAR-100: The symmetric noise setting is the same as that in MNIST and CIFAR-10. The asymmetric noise label is conducted following jocor: The 100 classes are grouped into

20 5-size superclasses, e.g. AQUATIC mammals contain BEAVER, DOLPHIN, OTTER, SEAL and WHALE. Within super-classes, each class is flipped into the next circularly with noise ratio τ.

Clothing1M: Clothing1M is a large-scale dataset from 14 clothing classes with noisy labels. The estimated noise level is roughly 40%.[193] we use the 1M images with noisy labels for training, the 14k and 10k clean data for validation and test, respectively. The 50k clean training data is not used since only noisy labels are required in training. We resize the image to $256 * 256$ and crop the middle $224 * 224$ as input, then perform normalization.

This dataset is seriously imbalanced among different classes and the label mistakes mostly happen between similar classes. Note that in PENCIL setup, they randomly selected a small balanced subset to relieve the difficulty caused by imbalance. The small subset includes 260k images and all classes have the same number of images. In our implementation, we use the raw imbalanced dataset which increase the learning difficulty.

Implementation details

We use a 2-layer MLP for MNIST, 7-layer CNN network architecture for CIFAR-10 and CIFAR-100 and 18-layer ResNet for Clothing1M, which are same in JoCoR for fair comparison. The network details are shown in Table. 3.2

For experiments on CIFAR-10 and CIFAR-100, Adam optimizer with momentum=0.9 is used and the batch size is 128. An initial learning rate of 0.001 is used in the first 250 epochs for iterative parameter and label learning. For fine-tuning the network with fixed labels, we run 50 epochs and linearly decay the learning rate from 0.001 to zero. The λ in Eq. 3.9 balances the supervised learning loss and agreement loss. We linearly decrease λ from 0.9 to 0.7 after each label update and keep 0.7 at the fine-tuning stage. The label update interval is 50 epochs. The label correction threshold $C_{retrain} = 5\%$.

In Clothing1M, the same Adam optimizer is used and the batch size is 64. We run 15 epochs in iterative optimization and 10 epochs for fine-tuning. The learning rate for the first 5 epochs is 8×10^{-4}, second 10 epochs is 5×10^{-4}, then for fine-tuning is 5×10^{-5}. λ is set

Table 3.2: The networks used on MNIST, CIFAR-10 and CIFAR-100

MLP on *MNIST*	CNN on *CIFAR-10* & *CIFAR-100*
28×28 Gray Image	32×32 RGB Image
Dense $28 \times 28 \longrightarrow 256$, ReLU	3×3, 64 BN, ReLU 3×3, 64 BN, ReLU 2×2 Max-pool
	3×3, 128 BN, ReLU 3×3, 128 BN, ReLU 2×2 Max-pool
	3×3, 196 BN, ReLU 3×3, 196 BN, ReLU 2×2 Max-pool
Dense $256 \longrightarrow 10$	Dense $256 \longrightarrow 100$

to 0.85. $T_{update} = 5$ and the retraining threshold $C_{retrain} = 5\%$.

Comparison with baselines

We compare our method with following baseline algorithms on each dataset: F-correction,[186] Co-teaching,[175] Co-teaching+,[194] JoCoR,[176] PENCIL,[178] DivideMix (Clothing1M dataset)[195] and L_{DMI}.[196] The performance of standard deep networks training on noisy datasets is also used as a simple baseline. The performance of PENCIL, DivideMix and L_{DMI} is based on our implementation, and the others are quoted from.[176] Since our work is inspired by JoCoR, the test accuracy comparison between JoCoR and ours during the last round of training is also presented.

MNIST: The test accuracy and standard deviation compared with baselines over MNIST are shown in Table. 3.3. In the flipping rate of symmetry-50%, symmetry-80% and asymmetry-40%, our method achieves the best performance among baselines. Besides our method, JoCoR, PENCIL and L_{DMI} performs better than other baselines in most cases. For Symmetry-20% noise, even there are 98% correct labels with our label update, JoCoR has a slightly higher mean accuracy than ours. JoCoR trains the network with filtered clean examples, it

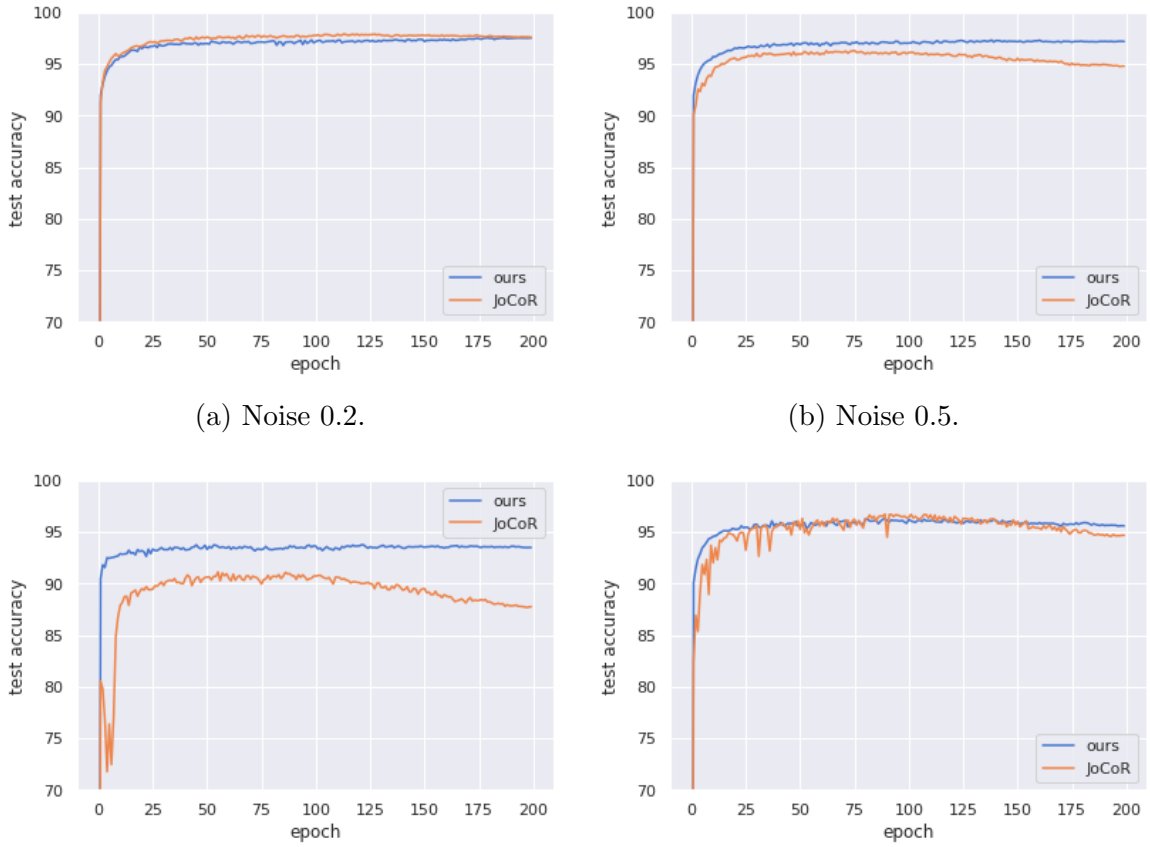
may because for this relatively easy classification task, a subset of clean data is enough to train a good classifier.

The MNIST column in Table. 3.4 shows the number of corrects labels in raw dataset and after label update. We corrected a large amount of noisy labels under each noise level. Besides improving the accuracy, the proposed iterative learning and correction also improves the model robustness. We can see as the noise ratio increased, the standard deviation is increasing which indicates the stability of models are affected by noise labels. Our method has a much smaller standard deviation especially in the high noise rate. For example, JoCoR and PENCIL have a relatively high performance in symmetric-80% noise and their standard deviation are 4.55 and 3.06, respectively. The proposed method has a much smaller standard deviation of 1.69. As shown in symmetric-80% noise, we have a small standard deviation compared to the other two highest performance methods: JoCoR and PENCIL. Therefore, our method has stronger robustness and more stable performance. The test accuracy during training compared with JoCoR is shown in Fig. 3.14

Table 3.3: Average test accuracy (%) and standard deviation on MNIST of 5 trials.

Flipping Rate	Standard	F-correction	Co-teaching	Co-teaching+	JoCoR	PENCIL	L_{DMI}	Ours
Symmetry-20%	79.56 ± 0.44	95.38 ± 0.10	95.10 ± 0.16	97.81 ± 0.03	**98.06 ± 0.04**	97.45±0.12	96.75	97.81 ± 0.03
Symmetry-50%	52.66 ± 0.43	92.74 ± 0.21	89.82 ± 0.31	95.80 ± 0.09	96.64 ± 0.12	96.89±0.68	95.21	**97.19 ± 0.32**
Symmetry-80%	23.43 ± 0.31	72.96 ± 0.90	79.73 ± 0.35	58.92 ± 14.73	84.89 ± 4.55	93.19± 3.06	91.98	**93.60 ± 1.69**
Asymmetry-40%	79.00 ± 0.28	89.77 ± 0.96	90.28 ± 0.27	93.28 ± 0.43	95.24 ± 0.10	95.07±0.28	96.02	**96.23 ± 0.11**

CIFAR-10: The test accuracy and standard deviation compared with baselines over CIFAR-10 are shown in Table. 3.5. Our method performs the best in all four cases especially with symmetric-50% noise (+6.21). We can see as the noise ratio increased, the standard deviation increases which indicates the stability of models is affected by noise labels. Our method has a smaller standard deviation especially in the high noise rate. Our standard deviation is 0.89 in symmetry-80%, while that of the other top-3 methods co-teaching, JoCoR and PENCIL are 2.22, 3.06 and 1.86, respectively. The test accuracy comparison with JoCoR



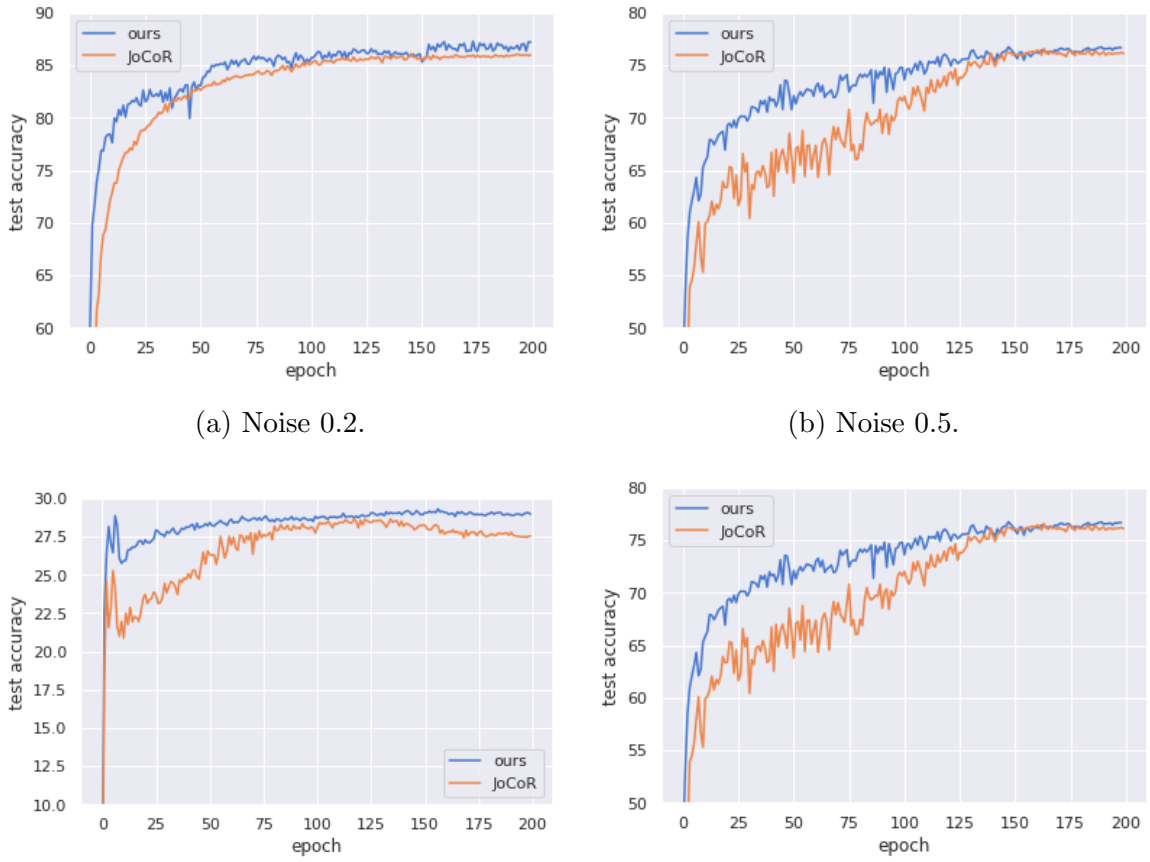
(a) Noise 0.2.

(b) Noise 0.5.

(c) Noise 0.8.

(d) Asymmetric noise 0.4.

Figure 3.14: Comparison with JoCoR on MNIST.



(a) Noise 0.2.

(b) Noise 0.5.

(c) Noise 0.8.

(d) Asymmetric noise 0.4.

Figure 3.15: Comparison with JoCoR on CIFAR-10.

is shown in Fig. 3.15

The CIFAR-10 column in Table. 3.4 shows the label correction results. We improve the clean label ratio from 80% and 50% to 91% and 86% in the first two symmetry noise. For the case of symmetry-80% and asymmetry-40%, a small portion of noisy labels are corrected, and this update also improves the performance compared to using joint training only in JoCoR.

CIFAR-100: we compare our results with other baselines on CIFAR-100 in Table 3.6. CIFAR-100 has a similar dataset size to CIFAR-10 while with 10 times of classes, thus the classification is more challenging. Our method is still the overall accuracy winner. In symmetry-20% and symmetry-50%, Our method and JoCoR works significantly better than other methods and ours has an advantage of $+1.6 \sim 1.9$ over JoCoR. In the hardest case

Table 3.4: Correct labels on MNIST, CIFAR-10 and CIFAR-100 under different label noise.

Dataset	MNIST		CIFAR-10		CIFAR-100	
	raw	update	raw	update	raw	update
Symmetry-20%	48021	58669	39962	45722	39976	41020
Symmetry-50%	30112	53904	25162	42974	25162	32220
Symmetry-80%	12097	45815	10000	11141	10000	12066
Asymmetry-40%	36109	56235	30115	31633	30115	32863

Table 3.5: Average test accuracy (%) and standard deviation on CIFAR-10 of 5 trials.

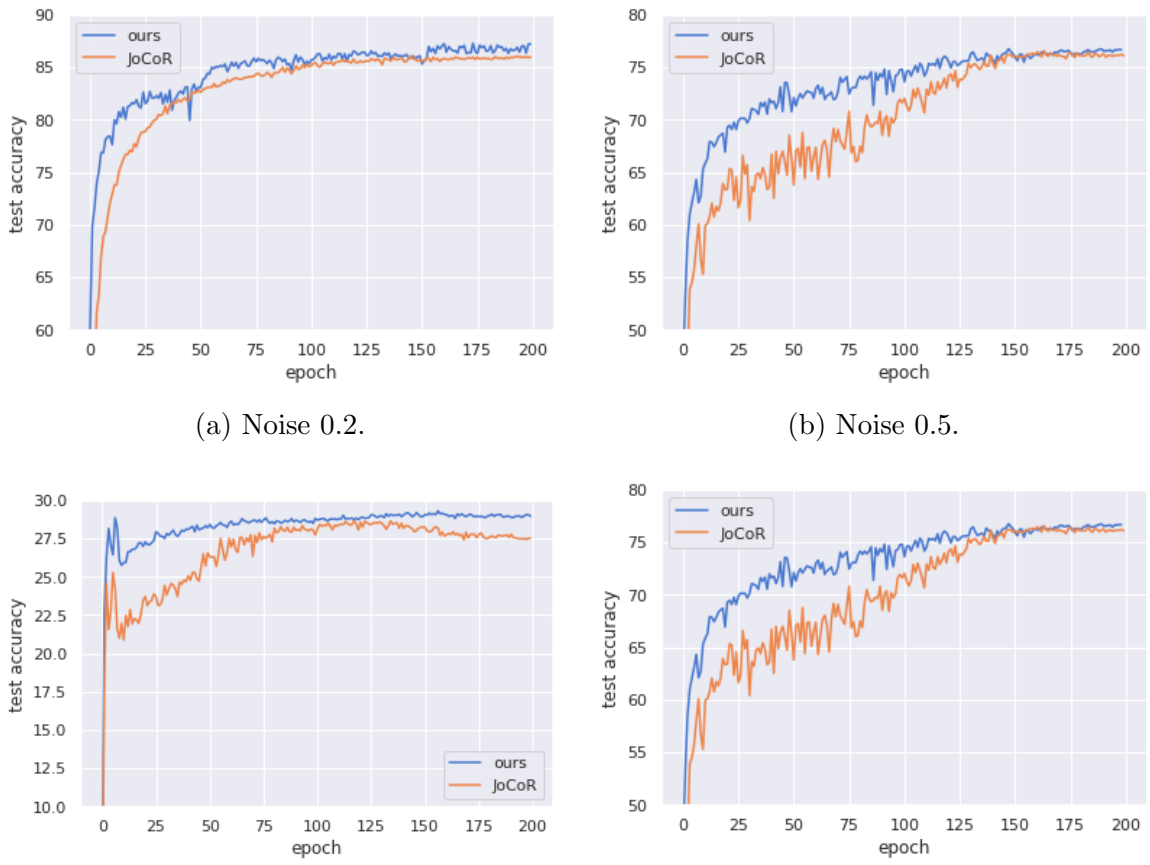
Flipping Rate	Standard	F-correction	Co-teaching	Co-teaching+	JoCoR	PENCIL	L_{DMI}	Ours
Symmetry-20%	69.18 ± 0.52	68.74 ± 0.20	78.23 ± 0.27	78.71 ± 0.34	85.73 ± 0.19	81.35 ± 0.32	80.52	**86.84 ± 0.18**
Symmetry-50%	42.71 ± 0.42	42.19 ± 0.60	71.30 ± 0.13	57.05 ± 0.54	79.41 ± 0.25	69.29 ±0.78	73.21	**85.62 ± 0.48**
Symmetry-80%	16.24 ± 0.39	15.88 ± 0.42	26.58 ± 2.22	24.19 ± 2.74	27.78 ± 3.06	25.30± 1.86	21.48	**29.01 ± 0.89**
Asymmetry-40%	69.43 ± 0.33	70.60 ± 0.40	73.78 ± 0.22	68.84 ± 0.20	76.36 ± 0.49	68.53±0.48	65.12	**76.83 ± 0.29**

of symmetry-80%, JoCoR ties together with Co-teaching and ours has +3.36 test accuracy. The standard classifier, F-correction and PENCIL fail in this case with the accuracy below 5%. In terms of asymmetry-40% noise, Co-teaching+ performs better than other baselines with 33.62% accuracy whereas that of ours is 34.41%. The test accuracy during training compared with JoCoR is shown in Fig. 3.17 The label correction under different cases are shown in CIFAR-100 column in Table. 3.4. The overall number of updated labels are not as large as in CIFAR-10 since there are not enough confident examples for the networks.

Table 3.6: Average test accuracy (%) and standard deviation on CIFAR-100 of 5 trials.

Flipping Rate	Standard	F-correction	Co-teaching	Co-teaching+	JoCoR	PENCIL	L_{DMI}	Ours
Symmetry-20%	35.14 ± 0.44	37.95 ± 0.10	43.73 ± 0.16	49.27 ± 0.03	53.01 ± 0.04	43.61 ±0.23	48.95	**54.90 ± 0.07**
Symmetry-50%	16.97 ± 0.40	24.98 ± 1.82	34.96 ± 0.50	40.04 ± 0.70	43.49 ± 0.46	26.41 ±0.51	39.21	**45.12 ± 0.42**
Symmetry-80%	4.41 ± 0.14	2.10 ± 2.23	15.15 ± 0.46	13.44 ± 0.37	15.49 ± 0.98	3.65±0.77	10.65	**18.85 ± 0.59**
Asymmetry-40%	27.29 ± 0.25	25.94 ± 0.44	28.35 ± 0.25	33.62 ± 0.39	32.70 ± 0.35	27.32 ±0.42	31.34	**34.41 ± 0.72**

Clothing1M: We demonstrate our method on real-world noisy labels with Clothing1M dataset, which includes a lot of unknown structure (asymmetric) noise. The results are shown in Table. 3.7, where "Best" denotes the test accuracy of the epoch where the validation



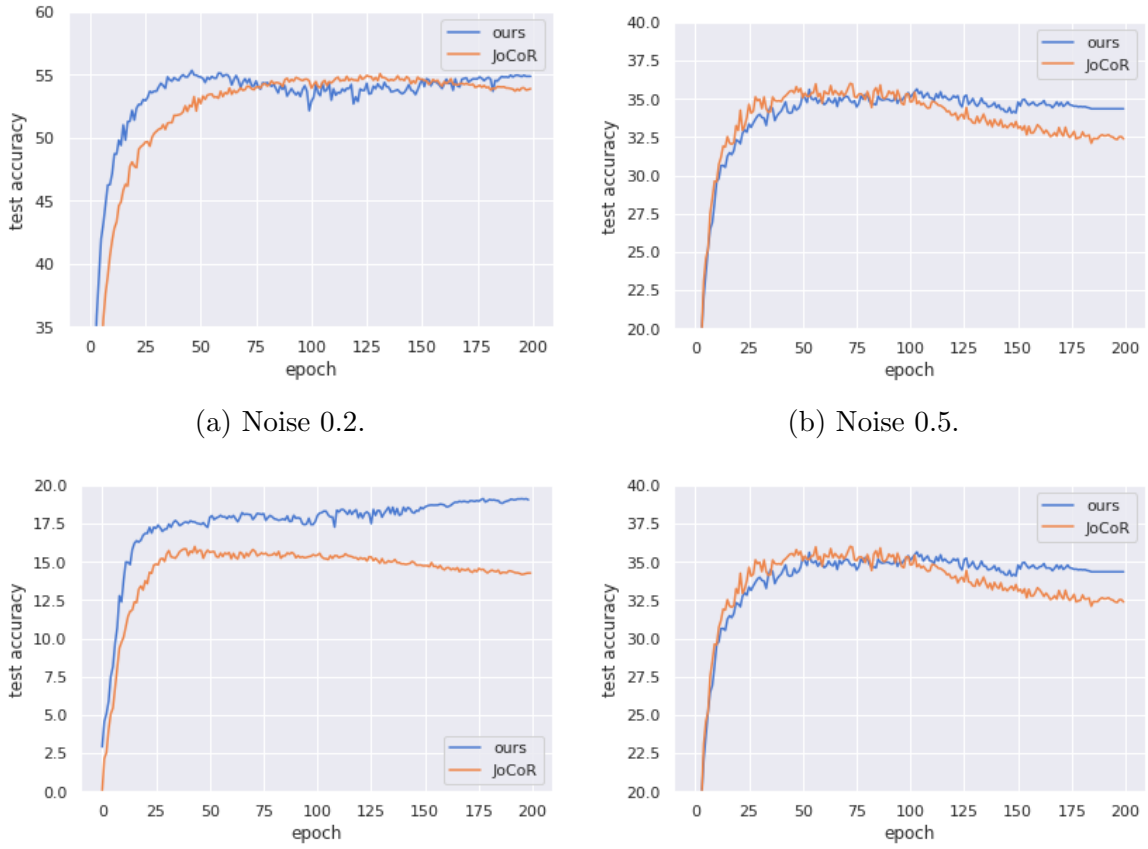
(a) Noise 0.2.

(b) Noise 0.5.

(c) Noise 0.8.

(d) Asymmetric noise 0.4.

Figure 3.16: Comparison with JoCoR on CIFAR-10.



(a) Noise 0.2.

(b) Noise 0.5.

(c) Noise 0.8.

(d) Asymmetric noise 0.4.

Figure 3.17: Comparison with JoCoR on CIFAR-100.

Table 3.7: Average accuracy (%) on Clothing1M test set.

Methods	Best	Last
Standard	67.22	64.68
F-correction[186]	68.93	65.36
Co-teaching[175]	69.21	68.51
Co-teaching+[194]	59.32	58.79
JoCoR[176]	70.30	69.79
PENCIL[178]	69.48	68.48
L_{DMI}[196]	70.79	**70.02**
DivideMix[195]	68.24	67.19
Ours	**71.63**	69.97

accuracy is optimal and "last" denotes the test accuracy of the last epoch. We use the complete noisy training set of Clothing1M instead of the small pseudo-balanced subset. Our method achieved SOTA performance in "Best". In the last epoch, our performance is slightly decreased because of overfitting and 0.05% lower than L_{DMI}. The accuracy drop can be improved by stopping training early according to the validation set in practice.

Ablation Study

We conduct ablation study for analyzing the effect of label update interval and retraining. The experiments are set up on CIFAR-10 with Symmetry-50% noise.

Label Update Interval: Instead of referring clean labels in each iteration,[177,178] our label correction is intermittent. With label update intervals, two classifiers can be trained on the current dataset with small loss selection strategy and try to maximize the use of current data. The agreement maximization principle will be more reliable with these relatively robust models. In this experiment, we set the label update interval to 50 epochs and compared it with the continuous update in each iteration. Similar to Joint optimization,[177] a pre-trained backbone is needed for continuous label update. We train two networks for 50 epochs before the label update both in continuous and intermittent methods.

Fig. 3.18 (a) and (b) show the test accuracy and number of correct labels versus epochs, respectively. In Fig. 3.18 (a), the large drop of the test accuracy indicates the trigger of retraining when the label correction rate is large. In continuous update, the label correction

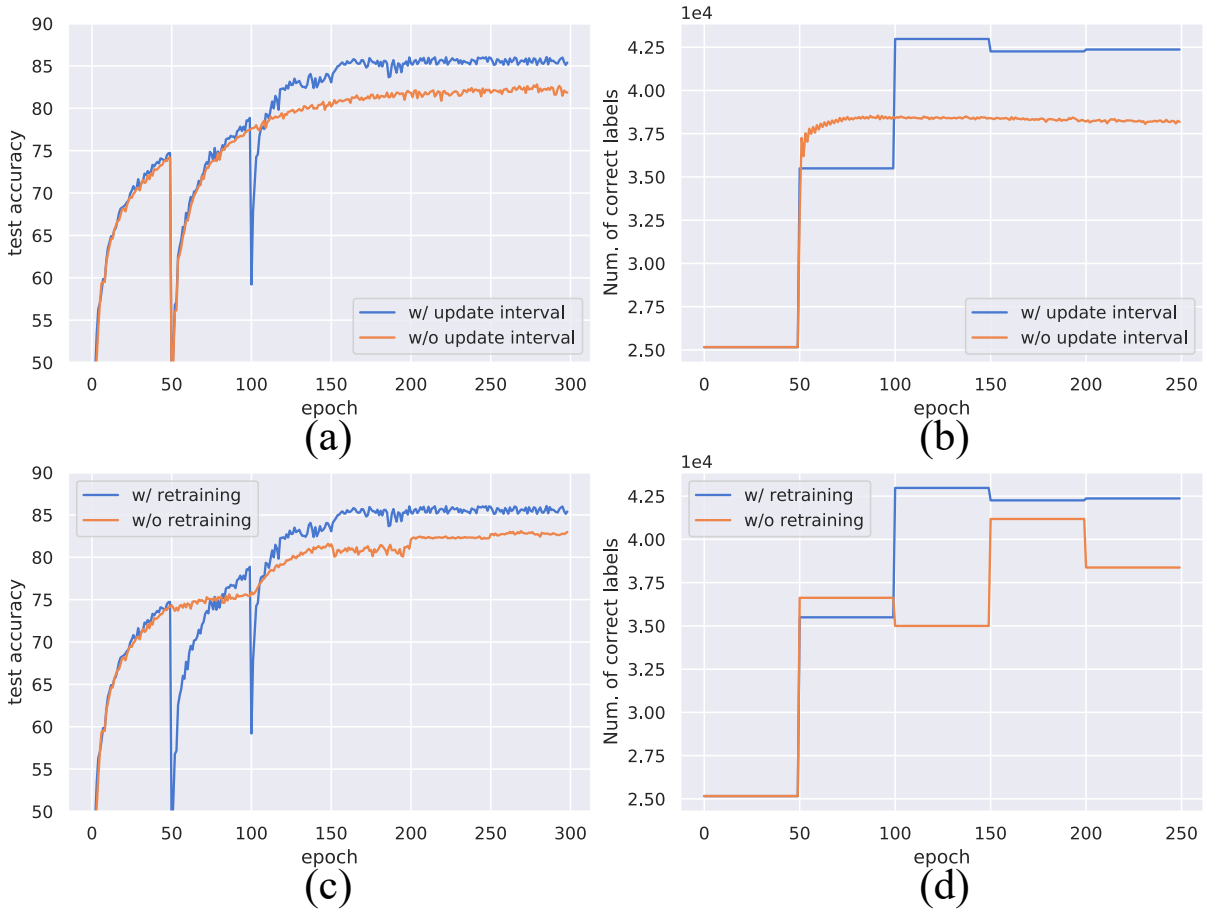


Figure 3.18: Comparison of ablation study over CIFAR-10. (a)(b): label update interval; (c)(d): retraining after correction.

is smooth after the first correction. For fair comparison, we retrain networks with new labels after the first update. We can see the test accuracy with update intervals is better than continuous update, and the number of correct labels is significantly higher than the latter. It proves the effectiveness of using relatively robust models to do the label correction. Note that continuous correction has an average test accuracy of 82.1%, which still maintains an advantage over the other methods in Table. 3.5.

Retraining After Correction Generally the network training with a fixed dataset is continuous for a large amount of epochs until it is stable. In Joint optimization and PENCIL[177,178] with label update, their labels and weights learning are performed in every iteration

and the dataset will not be changed dramatically after each label correction. In terms of our method, the labels may change a lot in every label update, because the correction ability of networks after the label correction interval can be improved much. In our experiment, retraining the models with the updated dataset has a better performance than continuous training as shown in Fig. 3.18 (c)(d). In Fig. 3.18 (d), the number of correct labels without retraining is less than retraining, and it drops a lot in the last update. Correspondingly, the test accuracy of the former is lower than the latter. However, the mean accuracy of 82.8% without retraining is still higher than that of other baselines.

3.2.5 Conclusion

We proposed a synergistic network learning and label correction methods to solve the noise label problem. We first jointly optimize the network parameters and noisy labels, then fine-tune the network with fixed labels. In the jointly optimization stage, we first train two stable networks on the current labels, then select the confident examples for label correction. Our method is independent of the backbone network thus it is easy to apply. We demonstrated our method with synthetic label noise on CIFAR-10 and CIFAR-100 and real-world large scale dataset Clothing1M with label noise. Our method is a winner compared with the baselines except on Clothing1M test set without stop the training early.

This framework can be directly applied to medical images with classification tasks. It can also extend to other tasks beyond image classification, such as object detection, segmentation, by replacing the loss function accordingly. The overall training strategy will remain the same.

Chapter 4

FUTURE WORK

4.1 Real time camera pose localization for tele-cystoscopy

4.1.1 Solving non-linear distribution of captured images in dictionary image retrieval

In the work of real time localization in bladder, we assume the scanning of robotic cystoscope will not have large non-linear movement, which is applicable in most cases. It will scan the full bladder layer by layer as shown in Fig. 2.23. Thus, in the current pipeline, we use the linear method PCA to perform the dimension reduction in the dictionary image retrieval and achieve a high success rate in the experiments over bladder phantoms. The PCA based dimension reduction can also be robust to a certain range of non-linear distributions as shown in Fig. 2.4. To make it more reliable in the real bladder test, which might have more non-linear movement of the robotic cystoscope caused by force of the bladder wall and folding of the cystoscope, the non-linear method for the dimension reduction can be explored in the future. The updated pipeline in the future can also be applied on other applications including several of our previous projects.

To start with, we can first add more non-linear movement when scanning the bladder phantom with the robotic cystoscope or use the clinical data collected manually which may not be as stable as the robotic platform and induce more non-linear movements.[139] The first direction of the exploration is to try other non-linear dimensional reduction approaches, such as kernal PCA, ISOMAP and diffusion map. It is approved that kernal PCA and ISOMAP does not over-perform PCA when there are no large non-linear changes of the data in the retina image application (2.3), and we will tweak the kernels and parameters of these non-linear approaches over the new dataset.

Another direction of solving the non-linear problem is to use neural network to learning

the embedding of images, which is a vector space, typically of lower dimension than the input space, which preserves relative dissimilarity (in the input space). An embedding is often unsupervised and constructed by transfer learning from large-scale annotated or non-annotated data.[197] Given an embedding, a downstream learning method, referred to as a two-stage method, can be applicable to the current data.[198,199] In our case, we can first build an image classification neural network trained on a labelled public dataset of endoscopic images and videos which contains similar features with the bladder environment.[200] Fig. 4.1 shows a simple example of the neural network structure for classification. During the training process, the network will learn the weights for both convolution layers and fully connected layers as shown in Fig. 4.1. The convolution layers become a feature extractor to extract the high level features of input images, and the fully connected layers before classification can learn the location relevance between the extracted features. After training, we can remove the top classification layer and the output of the previous layers will be a image representation feature vector, which has a much lower dimension than the original image. Normalization can be applied to reduce the sensitivity to illumination changes. When transferring to our bladder images captured with cystoscope, the feature vectors can be obtained directly with the trained network and the vectors form the dictionary in the first exam of the patient, which is similar with our current pipeline. With a new bladder frame in the following visits of the patient, the feature vector can be computed in the same way. We can use inner product as an unsupervised similarity metric to retrieve the dictionary image which has the largest inner product score with the current frame.

By training on the public endoscopic dataset,[200] we are able to fine-tune a feature vector to perform better within the particular distribution of the dataset. Even the endoscopic images in the public dataset has similarity with bladder images, training and testing within different domain distributions might still hurt performance. Compared to the unsupervised similarity metric, a supervised metric is assumed to perform better in the generalization. In our case, we can annotate a small set of nearest image pairs with our current method or collect them by controlling the robotic cystoscope. The two images are similar or not (1

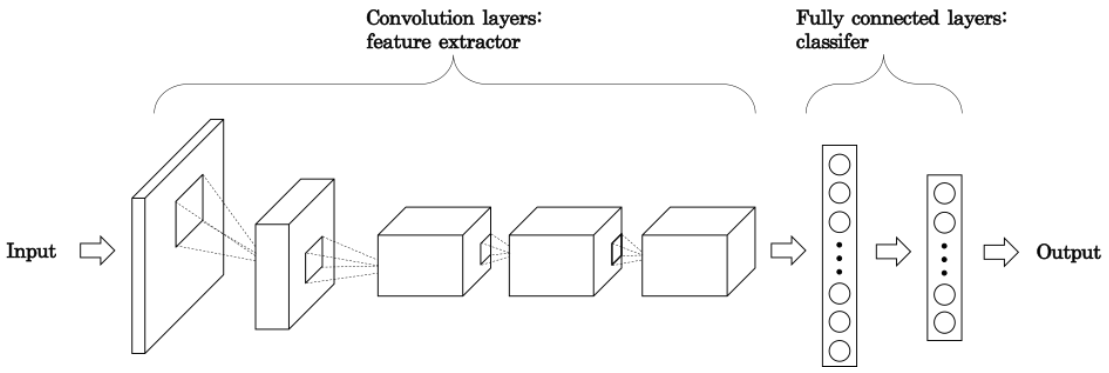


Figure 4.1: An example of the neural network structure for image classification.

and 0) can be determined by the threshold of the overlap area. According to the limit of our current 3D-to-2D correspondence, we can set the similarity to 0 when the overlap is too small to perform this step successfully. To learn the supervised similarity metric, weighted χ^2 similarity and the Siamese network can be used.[201]

4.1.2 Shape map for safe navigation

In the procedure of tele-cystoscopy, safety operation is critical. In our current pipeline, especially in the scanning of the bladder during patient first exam, the environment is fully unknown to our robotic cystoscope. On the other hand, we require high quality and fully covered videos in the first scanning to create the 3D model of the bladder, which means the cystoscope tip should be close to the bladder surface and we want to maintain an orthogonal as much as possible. In this case, a naive spiral path plan which is preset based on the general bladder shape (2.23) cannot avoid the collision with the bladder wall. Different bladders can have different shape after injecting liquid. It might cause unnecessary damage of the patient's bladder. To improve the safety of the scanning operation, an accuracy robot path planning for each patient is supposed to be created in advance for the following high-resolution robotic cystoscopy.

We plan to conduct a navigational scan to safely generate a shape map of the bladder for the path planning. The navigational scan is from the central region of the bladder to estimate general bladder shape and dimension safely, as shown in Fig. 4.2. Since the imaging distance

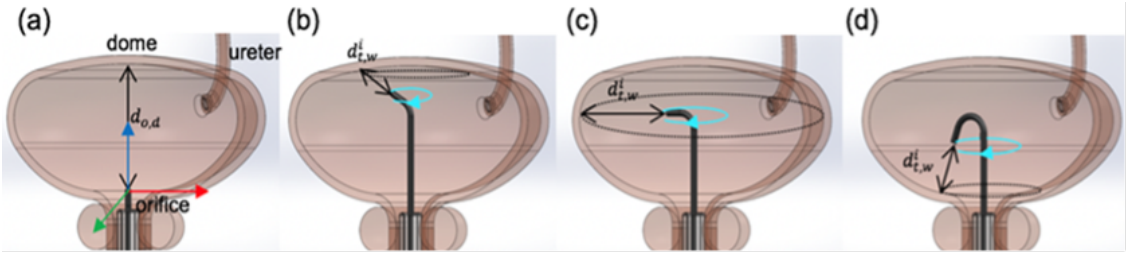


Figure 4.2: Initial navigational scan of the bladder once the scope has reached the orifice and the bladder is flushed for optical clarity and sizing (a).

is further, the captured video frames will not contain many fine-grained features like vessels but only the features related to the surface shape. To create the 3D navigational map with the frames with very sparse features, we plan to use deep learning for the depth estimation first and combined the robotic kinematics information for the 3D shape map generation from the depth map.

Deep learning methods have been widely used in estimating depth from monocular images with supervised learning.[202,203] Since large dataset with ground truth depth is hard to collect in medical applications, the unsupervised learning methods can also be used.[204,205] Another preferred option from the medical field is to generate a simulated training database of synthetic and virtual images with ground truth by developing a cystoscopic optical model. The depth map can be generated automatically for each frame. With the simulator we can create many training images with different settings and train a supervised depth estimation model. For the clinical data, we can transform it to the synthetic domain of the simulator with unsupervised domain adaptation, such as CycleGAN.[206] The depth map of the transformed image can be predicted by the trained model.[207]

Accuracy of the shape map generation process can be tested by placing electro-magnetic tracking sensors that fit alongside probes in the working channel of commercial flexible endoscopes.

4.2 End to end network in retina tracking

To improve the performance of the neural network and make it more robust to the scanning distortion, we perform the localization based on the ring level as the observation of Kalman filter. With a large amount of training dataset and tweaking the neural network properly, the ring level localization is supposed to be learned by the network. Since the outer ring area contains much fewer features compared with the full image, relying on the network to learn to focus on the outer ring regions by itself has a high requirement of the dataset size and model tweaking process. Another plan is to redesign the network to force its focus more on the features from outer rings, and it is convenient to be implemented in our current network structure. As shown in Fig. 3.7, when we get the feature map of the current frame F_{frame} and the full image F_{image}, F_{frame} is taken as the kernel to perform convolution on F_{image}. The values in F_{frame} represent the parameters of the convolution kernel. We can re-weight the importance of the kernel parameters by multiplying it with a weight mask: the outer ring area has the highest weights and the weights decrease to zero gradually as the radius shrinks. In this way, the outer rings of the frame can contribute more to the convolution with the full image, and have a larger influence on the prediction.

BIBLIOGRAPHY

[1] C. Gong, N. B. Erichson, J. P. Kelly, L. Trutoiu, B. T. Schowengerdt, S. L. Brunton, and E. J. Seibel, "Retinamatch: Efficient template matching of retina images for teleophthalmology," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1993–2004, 2019.

[2] J. Flandes, L. F. Giraldo-Cadavid, J. Alfayate, I. Fernández-Navamuel, C. Agusti, C. M. Lucena, A. Rosell, F. Andreo, C. Centeno, C. Montero *et al.*, "Bronchoscopist's perception of the quality of the single-use bronchoscope (ambu ascope4™) in selected bronchoscopies: a multicenter study in 21 spanish pulmonology services," *Respiratory research*, vol. 21, no. 1, pp. 1–9, 2020.

[3] C. Yin, L. Wei, K. Kose, A. K. Glaser, G. Peterson, M. Rajadhyaksha, and J. T. Liu, "Real-time video mosaicking to guide handheld in vivo microscopy," *Journal of Biophotonics*, p. e202000048, 2020.

[4] F. Altaf, S. M. Islam, N. Akhtar, and N. K. Janjua, "Going deep in medical image analysis: concepts, methods, challenges, and future directions," *IEEE Access*, vol. 7, pp. 99 540–99 572, 2019.

[5] "Build One – Vine Robots." [Online]. Available: https://www.vinerobots.org/build-one/

[6] M. W. Wintergerst, M. Petrak, J. Q. Li, P. P. Larsen, M. Berger, F. G. Holz, R. P. Finger, and T. U. Krohne, "Non-contact smartphone-based fundus imaging compared to conventional fundus imaging: a low-cost alternative for retinopathy of prematurity screening and documentation," *Scientific reports*, vol. 9, no. 1, pp. 1–8, 2019.

[7] X. Yao, T. Son, and J. Ma, "Developing portable widefield fundus camera for teleophthalmology: Technical challenges and potential solutions," *Experimental Biology and Medicine*, vol. 247, no. 4, pp. 289–299, 2022.

[8] M. J. Kundrat, P. G. Reinhall, and E. J. Seibel, "Method to achieve high frame rates in a scanning fiber endoscope," *Journal of Medical Devices*, vol. 5, no. 3, 2011.

[9] C. M. Lee, C. J. Engelbrecht, T. D. Soper, F. Helmchen, and E. J. Seibel, "Scanning fiber endoscopy with highly flexible, 1 mm catheterscopes for wide-field, full-color imaging," *Journal of biophotonics*, vol. 3, no. 5-6, pp. 385–407, 2010.

[10] C. Gong, S. Brunton, E. Seibel, L. Trutoiu, and B. Schowengerdt, "Real-time retinal localization for eye-tracking in head-mounted displays," in *Proceedings of Computer Vision and Pattern Recognition Conference (CVPR), Workshop on Computer Vision for AR/VR*, 2020.

[11] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," *Proceedings of the IEEE*, 2021.

[12] L. Oakden-Rayner, "Exploring large-scale public medical image datasets," *Academic radiology*, vol. 27, no. 1, pp. 106–112, 2020.

[13] ——, "Exploring the chestxray14 dataset: problems," *Wordpress: Luke Oakden Rayner*, 2017.

[14] S. Wang, K. Jin, H. Lu, C. Cheng, J. Ye, and D. Qian, "Human visual system-based fundus image quality assessment of portable fundus camera photographs," *IEEE transactions on medical imaging*, vol. 35, no. 4, pp. 1046–1055, 2016.

[15] L. Shi, H. Wu, J. Dong, K. Jiang, X. Lu, and J. Shi, "Telemedicine for detecting diabetic retinopathy: a systematic review and meta-analysis," *British Journal of Ophthalmology*, vol. 99, no. 6, pp. 823–831, 2015.

[16] I. N. Figueiredo, S. Kumar, C. M. Oliveira, J. D. Ramos, and B. Engquist, "Automated lesion detectors in retinal fundus images," *Computers in biology and medicine*, vol. 66, pp. 47–65, 2015.

[17] R. Kawasaki, N. Cheung, J. J. Wang, R. Klein, B. E. Klein, M. F. Cotch, A. R. Sharrett, S. Shea, F. A. Islam, and T. Y. Wong, "Retinal vessel diameters and risk of hypertension: the Multiethnic Study of Atherosclerosis." *Journal of hypertension*, vol. 27, no. 12, pp. 2386–93, 2009.

[18] N. Panwar, P. Huang, J. Lee, P. A. Keane, T. S. Chuan, A. Richhariya, S. Teoh, T. H. Lim, and R. Agrawal, "Fundus photography in the 21st century—a review of recent technological advances and their implications for worldwide healthcare," *Telemedicine and e-Health*, vol. 22, no. 3, pp. 198–208, 2016.

[19] W. Fink, M. A. Tarbell, and K. Garcia, "Smart ophthalmics: The future in teleophthalmology has arrived," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 9836, 2016.

[20] K. Roesch, T. Swedish, and R. Raskar, "Automated retinal imaging and trend analysis–a tool for health monitoring," *Clinical ophthalmology (Auckland, NZ)*, vol. 11, p. 1015, 2017.

[21] C. A. Ludwig, "The future of automated mobile eye diagnosis."

[22] H. Yu, E. S. Barriga, C. Agurto, S. Echegaray, M. S. Pattichis, W. Bauman, and P. Soliz, "Fast localization and segmentation of optic disk in retinal images using directional matched filtering and level sets," *IEEE Transactions on information technology in biomedicine*, vol. 16, no. 4, pp. 644–657, 2012.

[23] X. Zhang, G. Thibault, E. Decencière, B. Marcotegui, B. Laÿ, R. Danno, G. Cazuguel, G. Quellec, M. Lamard, P. Massin *et al.*, "Exudate detection in color retinal images for mass screening of diabetic retinopathy," *Medical image analysis*, vol. 18, no. 7, pp. 1026–1043, 2014.

[24] A. D. Mora, J. Soares, and J. M. Fonseca, "A template matching technique for artifacts detection in retinal images," in *Image and Signal Processing and Analysis (ISPA), 2013 8th International Symposium on*. IEEE, 2013, pp. 717–722.

[25] C. V. Stewart, C.-L. Tsai, and B. Roysam, "The dual-bootstrap iterative closest point algorithm with application to retinal image registration," *IEEE transactions on medical imaging*, vol. 22, no. 11, pp. 1379–1394, 2003.

[26] M. Sofka and C. V. Stewart, "Retinal vessel centerline extraction using multiscale matched filters, confidence and edge measures," *IEEE transactions on medical imaging*, vol. 25, no. 12, pp. 1531–1546, 2006.

[27] J. Xu, O. Chutatape, E. Sung, C. Zheng, and P. C. T. Kuan, "Optic disk feature extraction via modified deformable model technique for glaucoma analysis," *Pattern recognition*, vol. 40, no. 7, pp. 2063–2076, 2007.

[28] I. N. Figueiredo, S. Moura, J. S. Neves, L. Pinto, S. Kumar, C. M. Oliveira, and J. D. Ramos, "Automated retina identification based on multiscale elastic registration," *Computers in biology and medicine*, vol. 79, pp. 130–143, 2016.

[29] I. N. Figueiredo, J. S. Neves, S. Moura, C. M. Oliveira, and J. D. Ramos, "Pattern classes in retinal fundus images based on function norms," in *International Symposium Computational Modeling of Objects Represented in Images*. Springer, 2014, pp. 95–105.

[30] Y. Wang, J. Shen, W. Liao, and L. Zhou, "Automatic fundus images mosaic based on sift feature," in *Image and Signal Processing (CISP), 2010 3rd International Congress on*, vol. 6. IEEE, 2010, pp. 2747–2751.

[31] C.-L. Tsai, C.-Y. Li, G. Yang, and K.-S. Lin, "The edge-driven dual-bootstrap iterative closest point algorithm for registration of multimodal fluorescein angiogram sequence," *IEEE transactions on medical imaging*, vol. 29, no. 3, pp. 636–649, 2010.

[32] G. Wang, Z. Wang, Y. Chen, and W. Zhao, "Robust point matching method for multimodal retinal image registration," *Biomedical Signal Processing and Control*, vol. 19, pp. 68–76, 2015.

[33] C. Hernandez-Matas, X. Zabulis, and A. A. Argyros, "Retinal image registration based on keypoint correspondences, spherical eye modeling and camera pose estimation," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 5650–5654.

[34] K. J. Friston, J. Ashburner, C. D. Frith, J.-B. Poline, J. D. Heather, and R. S. Frackowiak, "Spatial registration and normalization of images," *Human brain mapping*, vol. 3, no. 3, pp. 165–189, 1995.

[35] A. V. Cideciyan, "Registration of ocular fundus images: an algorithm using cross-correlation of triple invariant image descriptors," *IEEE Engineering in Medicine and Biology Magazine*, vol. 14, no. 1, pp. 52–58, 1995.

[36] Y.-M. Zhu, "Mutual information-based registration of temporal and stereo retinal images using constrained optimization," *Computer methods and programs in biomedicine*, vol. 86, no. 3, pp. 210–215, 2007.

[37] E. Liberty, F. Woolfe, P.-G. Martinsson, V. Rokhlin, and M. Tygert, "Randomized algorithms for the low-rank approximation of matrices," *Proceedings of the National Academy of Sciences*, vol. 104, no. 51, pp. 20 167–20 172, 2007.

[38] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM review*, vol. 53, no. 2, pp. 217–288, 2011.

[39] N. B. Erichson, A. Mendible, S. Wihlborn, and J. N. Kutz, "Randomized nonnegative matrix factorization," *Pattern Recognition Letters*, vol. 104, pp. 1–7, 2018.

[40] "Structured analysis of the retina," *http://www.ces.clemson.edu /˜ahoover/stare/*, accessed on May 15, 2018.

[41] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[42] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[43] J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor, *Dynamic mode decomposition: data-driven modeling of complex systems*. SIAM, 2016, vol. 149.

[44] N. B. Erichson, S. Voronin, S. L. Brunton, and J. N. Kutz, "Randomized matrix decompositions using r," *arXiv preprint arXiv:1608.02148*, 2016.

[45] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust Principal Component Analysis?" *Journal of the ACM*, vol. 58, no. 3, pp. 11–37, 2011.

[46] T. Bouwmans, S. Javed, H. Zhang, Z. Lin, and R. Otazo, "On the applications of robust pca in image and video processing," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1427–1457, 2018.

[47] A. Baghaie, R. M. D'souza, and Z. Yu, "Sparse and low rank decomposition based batch image alignment for speckle reduction of retinal oct images," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2015, pp. 226–230.

[48] Y. Fu, C. Wang, Y. Wang, B. Chen, Q. Peng, and L. Wang, "Automatic detection of longitudinal changes for retinal fundus images based on low-rank decomposition," *Journal of Medical Imaging and Health Informatics*, vol. 8, no. 2, pp. 284–294, 2018.

[49] P. Thevenaz and M. A. Unser, "Spline pyramids for intermodal image registration using mutual information," in *Wavelet Applications in Signal and Image Processing V*, vol. 3169. International Society for Optics and Photonics, 1997, pp. 236–247.

[50] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellen, and W. Eubank, "Pet-ct image registration in the chest using free-form deformations," *IEEE transactions on medical imaging*, vol. 22, no. 1, pp. 120–128, 2003.

[51] J.-M. Morel and G. Yu, "Asift: A new framework for fully affine invariant image comparison," *SIAM journal on imaging sciences*, vol. 2, no. 2, pp. 438–469, 2009.

[52] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *International conference on artificial neural networks*. Springer, 1997, pp. 583–588.

[53] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[54] L. J. van Rijn, "Torsional eye movements in humans," 1994.

[55] T. Walter, J.-C. Klein, P. Massin, and A. Erginay, "A contribution of image processing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus images of the human retina," *IEEE transactions on medical imaging*, vol. 21, no. 10, pp. 1236–1243, 2002.

[56] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[57] J. M. Fitzpatrick, J. B. West, and C. R. Maurer, "Predicting error in rigid-body point-based registration," *IEEE transactions on medical imaging*, vol. 17, no. 5, pp. 694–702, 1998.

[58] X. Dai and M. Bikdash, "Trend analysis of fragmented time series for mhealth apps: Hypothesis testing based adaptive spline filtering method with importance weighting," *IEEE Access*, vol. 5, pp. 27 767–27 776, 2017.

[59] C. Gong, J. P. Kelly, L. Trutoiu, B. Schowengerdt, S. Brunton, and E. Seibel, "Measurement of retinal vessel width in tele-ophthalmology for mobile health monitoring," *Investigative Ophthalmology & Visual Science*, vol. 59, no. 9, pp. 4624–4624, 2018.

[60] X. Xu, W. Ding, X. Wang, R. Cao, M. Zhang, P. Lv, and F. Xu, "Smartphone-based accurate analysis of retinal vasculature towards point-of-care diagnostics," *Scientific reports*, vol. 6, p. 34603, 2016.

[61] I. Yusuf, J. Barnes, T. Fung, J. Elston, and C. Patel, "Non-contact ultra-widefield retinal imaging of infants with suspected abusive head trauma," *Eye*, vol. 31, no. 3, p. 353, 2017.

[62] D. Hu, Y. Gong, E. J. Seibel, L. N. Sekhar, and B. Hannaford, "Semi-autonomous image-guided brain tumour resection using an integrated robotic system: A bench-top study," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 14, no. 1, p. e1872, 2018.

[63] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 59–73, 2007.

[64] T. Bergen and T. Wittenberg, "Stitching and surface reconstruction from endoscopic image sequences: a review of applications and methods," *IEEE journal of biomedical and health informatics*, vol. 20, no. 1, pp. 304–321, 2014.

[65] K. E. Loewke, D. B. Camarillo, W. Piyawattanametha, M. J. Mandella, C. H. Contag, S. Thrun, and J. K. Salisbury, "In vivo micro-image mosaicing," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 159–171, 2010.

[66] A. Bontala, J. Sivaswamy, and R. R. Pappuru, "Image mosaicing of low quality neonatal retinal images," in *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2012, pp. 720–723.

[67] J. Jalili, S. M. Hejazi, M. Riazi-Esfahani, A. Eliasi, M. Ebrahimi, M. Seydi, M. A. Fard, and A. Ahmadian, "Retinal image mosaicking using scale-invariant feature transformation feature descriptors and voronoi diagram," *Journal of Medical Imaging*, vol. 7, no. 4, p. 044001, 2020.

[68] R. Szeliski and H.-Y. Shum, "Creating full view panoramic image mosaics and environment maps," in *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, 1997, pp. 251–258.

[69] K. Kose, M. Gou, O. Yélamos, M. Cordova, A. M. Rossi, K. S. Nehal, E. S. Flores, O. Camps, J. G. Dy, D. H. Brooks *et al.*, "Automated video-mosaicking approach for confocal microscopic imaging : an approach to address challenges in imaging living tissue and extend field of view," *Scientific reports*, vol. 7, no. 1, pp. 1–11, 2017.

[70] J. Westwood *et al.*, "Real-time image mosaicing for medical applications," *Medicine Meets Virtual Reality 15: In Vivo, in Vitro, in Silico: Designing the Next in Medicine*, vol. 125, p. 304, 2007.

[71] Y. S. Chang, E. di Tomaso, D. M. McDonald, R. Jones, R. K. Jain, and L. L. Munn, "Mosaic blood vessels in tumors: frequency of cancer cells in contact with flowing blood," *Proceedings of the National Academy of Sciences*, vol. 97, no. 26, pp. 14 608–14 613, 2000.

[72] R. Szeliski *et al.*, "Image alignment and stitching: A tutorial," *Foundations and Trends® in Computer Graphics and Vision*, vol. 2, no. 1, pp. 1–104, 2007.

[73] V. Kwatra, A. Schödl, I. Essa, G. Turk, and A. Bobick, "Graphcut textures: image and video synthesis using graph cuts," *ACM Transactions on Graphics (ToG)*, vol. 22, no. 3, pp. 277–286, 2003.

[74] A. Eden, M. Uyttendaele, and R. Szeliski, "Seamless image stitching of scenes with large motions and exposure differences," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 2498–2505.

[75] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *ACM SIGGRAPH 2003 Papers*, 2003, pp. 313–318.

[76] S. Ghannam and A. L. Abbott, "Cross correlation versus mutual information for image mosaicing," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 4, no. 11, 2013.

[77] N. Gracias and J. Santos-Victor, "Underwater video mosaics as visual navigation maps," *Computer Vision and Image Understanding*, vol. 79, no. 1, pp. 66–91, 2000.

[78] A. Dame and E. Marchand, "Video mosaicing using a mutual information-based motion estimation process," in *2011 18th IEEE International Conference on Image Processing*, 2011, pp. 1493–1496.

[79] R. Miranda-Luna, C. Daul, W. C. P. M. Blondel, Y. Hernandez-Mier, D. Wolf, and F. Guillemin, "Mosaicing of bladder endoscopic image sequences: Distortion calibration and registration algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 2, pp. 541–553, 2008.

[80] A. Dame and E. Marchand, "Second-order optimization of mutual information for real-time image registration," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 4190–4203, 2012.

[81] Y. Hernandez-Mier, W. Blondel, C. Daul, D. Wolf, and F. Guillemin, "Fast construction of panoramic images for cystoscopic exploration," *Computerized Medical Imaging and Graphics*, vol. 34, no. 7, pp. 579–592, 2010.

[82] A. B. Hamadou, C. Soussen, W. Blondel, C. Daul, and D. Wolf, "Comparative study of image registration techniques for bladder video-endoscopy," in *European Conference on Biomedical Optics*. Optical Society of America, 2009, p. 7371_18.

[83] R. Miranda-Luna, C. Daul, W. C. Blondel, Y. Hernandez-Mier, D. Wolf, and F. Guillemin, "Mosaicing of bladder endoscopic image sequences: Distortion calibration and registration algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 2, pp. 541–553, 2008.

[84] R. Miranda-Luna, Y. Hernandez-Mier, C. Daul, W. C. Blondel, and D. Wolf, "Mosaicing of medical video-endoscopic images: data quality improvement and algorithm testing," in *(ICEEE). 1st International Conference on Electrical and Electronics Engineering, 2004.* IEEE, 2004, pp. 530–535.

[85] A. Kumar, R. S. Bandaru, B. M. Rao, S. Kulkarni, and N. Ghatpande, "Automatic image alignment and stitching of medical images with seam blending," *World Academy of Science, Engineering and Technology*, vol. 65, 2012.

[86] S. L. Brunton and J. N. Kutz, *Singular Value Decomposition (SVD)*. Cambridge University Press, 2019, p. 3–46.

[87] E. Liberty, F. Woolfe, P.-G. Martinsson, V. Rokhlin, and M. Tygert, "Randomized algorithms for the low-rank approximation of matrices," *Proceedings of the National Academy of Sciences*, vol. 104, no. 51, pp. 20 167–20 172, 2007.

[88] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM review*, vol. 53, no. 2, pp. 217–288, 2011.

[89] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 1–37, 2011.

[90] G. Peterson, D. K. Zanoni, M. Ardigo, J. C. Migliacci, S. G. Patel, and M. Rajadhyaksha, "Feasibility of a video-mosaicking approach to extend the field-of-view for reflectance confocal microscopy in the oral cavity in vivo," *Lasers in Surgery and Medicine*, vol. 51, no. 5, pp. 439–451, 2019.

[91] O. Yélamos, B. P. Hibler, M. Cordova, T. J. Hollmann, K. Kose, M. A. Marchetti, P. L. Myskowski, M. P. Pulitzer, M. Rajadhyaksha, A. M. Rossi *et al.*, "Handheld reflectance confocal microscopy for the detection of recurrent extramammary paget disease," *JAMA dermatology*, vol. 153, no. 7, pp. 689–693, 2017.

[92] E. Flores, O. Yélamos, M. Cordova, K. Kose, W. Phillips, E. Lee, A. Rossi, K. Nehal, and M. Rajadhyaksha, "Peri-operative delineation of non-melanoma skin cancer margins in vivo with handheld reflectance confocal microscopy and video-mosaicking," *Journal of the European Academy of Dermatology and Venereology*, vol. 33, no. 6, pp. 1084–1091, 2019.

[93] L. Wei, C. Yin, Y. Fujita, N. Sanai, and J. T. Liu, "Handheld line-scanned dual-axis confocal microscope with pistoned mems actuation for flat-field fluorescence imaging," *Optics letters*, vol. 44, no. 3, pp. 671–674, 2019.

[94] J. Engelsgjerd and C. Deibert, "Cystoscopy," Jan 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK493180/?report=classic

[95] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer Statistics, 2021." *CA: a cancer journal for clinicians*, vol. 71, no. 1, pp. 7–33, jan 2021.

[96] K. Chamie, M. S. Litwin, J. C. Bassett, T. J. Daskivitch, J. Lai, J. M. Hanley, B. R. Konety, C. S. Saigal, and T. U. D. i. A. Project, "Recurrence of high-risk bladder cancer: a population-based analysis," *Cancer*, vol. 119, no. 17, pp. 3219–3227, 2014.

[97] National Comprehensive Cancer Network, "Bladder Cancer (Version 4.2021)," url = https://www.nccn.org /professionals/physician_gls/pdf/bladder.pdf.

[98] "The State of the Urology Workforce and Practice in the United States," American Urology Association, Tech. Rep., 2019. [Online]. Available: https://www.auanet.org/research/research-resources/aua-census/census-overview

[99] K. L. Lurie, R. Angst, D. V. Zlatev, J. C. Liao, and A. K. Ellerbee Bowden, "3d reconstruction of cystoscopy videos for comprehensive bladder records," *Biomedical optics express*, vol. 8, no. 4, 2017.

[100] "Chapter 4 - endoscopic treatment of bladder tumors," in *Endoscopic Diagnosis and Treatment in Urinary Bladder Pathology*, P. A. Geavlete, Ed. San Diego: Academic Press, 2016, pp. 83–203.

[101] M. Brown, "Autostitch," 2018, http://matthewalunbrown.com/autostitch/autostitch.html.

[102] M. Mossanen and J. L. Gore, "The burden of bladder cancer care: direct and indirect costs." *Current opinion in urology*, vol. 24, no. 5, pp. 487–491, sep 2014.

[103] B. K. Hollenbeck, R. L. Dunn, Z. Ye, and J. M. Hollingsworth, "Delays in Diagnosis and Bladder Cancer Mortality," *Cancer*, pp. 5235–5242, 2010.

[104] J. Marescaux and F. Rubino, "Transcontinental robot-assisted remote telesurgery, feasibility and potential applications," *Annals of Surgery*, vol. 235, no. 4, pp. 487–492, 2006.

[105] C. F. Graetzel, A. Sheehy, and D. P. Noonan, "Robotic bronchoscopy drive mode of the Auris Monarch platform," in *International Conference on Robotics and Automation (ICRA)*. Montreal, Montreal, Canada: IEEE, 2019, pp. 3895–3901.

[106] L. Yarmus, J. Akulian, M. Wahidi, A. Chen, J. P. Steltz, S. L. Solomon, D. Yu, F. Maldonado, J. Cardenas-Garcia, D. Molena, H. Lee, and A. Vachani, "A Prospective Randomized Comparative Study of Three Guided Bronchoscopic

Approaches for Investigating Pulmonary Nodules: The PRECISION-1 Study," in *Chest*, vol. 157, no. 3. Elsevier Inc, mar 2020, pp. 694–701. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31678307/

[107] N. Sarli, G. Del Giudice, S. De, M. S. DIetrich, S. D. Herrell, and N. Simaan, "Preliminary Porcine in Vivo Evaluation of a Telerobotic System for Transurethral Bladder Tumor Resection and Surveillance," *Journal of Endourology*, vol. 32, no. 6, pp. 516–522, 2018.

[108] N. Sarli, G. Del Giudice, S. De, M. S. Dietrich, S. D. Herrell, and N. Simaan, "TURBot: A system for robot-assisted transurethral bladder tumor resection," *IEEE/ASME Transactions on Mechatronics*, vol. 24, no. 4, pp. 1452–1463, 2020.

[109] R. J. Hendrick, C. R. Mitchell, S. Duke Herrell, and R. J. W. Iii, "Hand-held transendoscopic robotic manipulators: A transurethral laser prostate surgery case study," *The International Journal of Robotics Research*, vol. 34, no. 13, pp. 1559–1572, 2015.

[110] A. Wong, Y. Phan, H. Thursby, and W. Mahmalji, "The First UK Experience with Single-use Disposable Flexible Cystoscopes: An In-depth Cost Analysis, Service Delivery and Patient Satisfaction Rate with Ambu® aScope™ 4 Cysto," *Journal of Endoluminal Endourology*, vol. 4, no. 1, pp. e29–e44, apr 2021. [Online]. Available: http://www.jeleu.com/index.php/JELEU/article/view/120

[111] A. Schwein, B. Kramer, P. Chinnadurai, S. Walker, M. O'Malley, A. Lumsden, and J. Bismuth, "Flexible robotics with electromagnetic tracking improves safety and efficiency during in vitro endovascular navigation," *Journal of Vascular Surgery*, vol. 65, no. 2, pp. 530–537, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.jvs.2016.01.045

[112] E. D. Rozeboom, R. Reilink, M. P. Schwartz, P. Fockens, and I. A. M. J. Broeders, "Evaluation of the tip-bending response in clinically used endoscopes," *Int J Med Robotics Comput Assist Surg*, vol. 9, pp. 240–246, 2013. [Online]. Available: http://dx.doi.org/

[113] B. Bardou, F. Nageotte, P. Zanne, and M. De Mathelin, "Improvements in the control of a flexible endoscopic system," in *Proceedings - IEEE International Conference on Robotics and Automation*. Institute of Electrical and Electronics Engineers Inc., 2012, pp. 3725–3732.

[114] L. Sliker, G. Ciuti, M. Rentschler, and A. Menciassi, "Magnetically driven medical devices: a review," *Expert review of medical devices*, vol. 12, no. 6, pp. 737–752, 2015.

[115] J. Li, E. S. Barjuei, G. Ciuti, Y. Hao, P. Zhang, A. Menciassi, Q. Huang, and P. Dario, "Magnetically-driven medical robots: An analytical magnetic model for endoscopic capsules design," *Journal of Magnetism and Magnetic Materials*, vol. 452, pp. 278–287, 2018.

[116] F. Bianchi, A. Masaracchia, E. Shojaei Barjuei, A. Menciassi, A. Arezzo, A. Koulaouzidis, D. Stoyanov, P. Dario, and G. Ciuti, "Localization strategies for robotic endoscopic capsules: a review," *Expert review of medical devices*, vol. 16, no. 5, pp. 381–403, 2019.

[117] C. Fang, W. Sang, J. D. J. Gumprecht, G. Strauss, and T. C. Lueth, "Image-guided steering of a motorized hand-held flexible rhino endoscope in ENT diagnoses," *2012 IEEE International Conference on Robotics and Biomimetics, ROBIO 2012 - Conference Digest*, pp. 1086–1091, 2012.

[118] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[119] O. G. Grasa, E. Bernal, S. Casado, I. Gil, and J. M. M. Montiel, "Visual slam for handheld monocular endoscope," *IEEE Transactions on Medical Imaging*, vol. 33, no. 1, pp. 135–146, 2014.

[120] N. Mahmoud, T. Collins, A. Hostettler, L. Soler, C. Doignon, and J. M. M. Montiel, "Live tracking and dense reconstruction for handheld monocular endoscopy," *IEEE Transactions on Medical Imaging*, vol. 38, no. 1, pp. 79–89, 2019.

[121] C. Xie, T. Yao, J. Wang, and Q. Liu, "Endoscope localization and gastrointestinal feature map construction based on monocular slam technology," *Journal of Infection and Public Health*, vol. 13, no. 9, pp. 1314–1321, 2020.

[122] ——, "Endoscope localization and gastrointestinal feature map construction based on monocular slam technology," *Journal of infection and public health*, vol. 13, no. 9, pp. 1314–1321, 2020.

[123] T. D. Soper, M. P. Porter, and E. J. Seibel, "Surface mosaics of the bladder reconstructed from endoscopic video for automated surveillance," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 6, pp. 1670–1680, 2012.

[124] A. Ben-Hamadou, C. Daul, and C. Soussen, "Construction of extended 3d field of views of the internal bladder wall surface: a proof of concept," *CoRR*, vol. abs/1607.04773, 2016.

[125] Q. e. a. Péntek, "Image-based 3d surface approximation of the bladder using structure-from-motion for enhanced cystoscopy based on phantom data," *Biomedizinische Technik. Biomedical engineering*, vol. 63, no. 4, 2018.

[126] N. O. Falcon, S. Ranjbar, E. Cisneros, B. Vu, A. Schoppe, P. Sanchez, Y. Jin, J. Ye, Y. Feng, D. Kaushik, and R. L. Hood, "Innovative computer vision approach to 3D bladder model reconstruction from flexible cystoscopy," in *Therapeutics and Diagnostics in Urology 2019*, vol. 10852. SPIE, 2019, pp. 18 – 26.

[127] Y. Zhou, R. L. Eimen, E. J. Seibel, and A. K. Bowden, "Cost-efficient video synthesis and evaluation for development of virtual 3d endoscopy," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, p. 1800711, 2021.

[128] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[129] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[130] A. S. Vemuri, K.-C. Liu, Y. Ho, H.-S. Wu, and M.-C. Ku, "Endoscopic video mosaicing: application to surgery and diagnostics," in *Living imaging workshop*, 2011, pp. 1–2.

[131] D. K. Iakovidis, E. Spyrou, and D. Diamantis, "Efficient homography-based video visualization for wireless capsule endoscopy," in *13th IEEE International Conference on BioInformatics and BioEngineering*. IEEE, 2013, pp. 1–4.

[132] R. Richa, B. Vágvölgyi, M. Balicki, G. Hager, and R. H. Taylor, "Hybrid tracking and mosaicking for information augmentation in retinal surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2012, pp. 397–404.

[133] "Cancer stat facts: Bladder Cancer (National Cancer Institute, Surveillance, Epidemiology, and End Results (SEER) Program), 2018."

[134] C. Wengert and M. Reeff, "Fully automatic endoscope calibration for intraoperative use," in *in [Bildverarbeitung in der Medizin], Informatik aktuell, 419–423, Springer Berlin*, 2006.

[135] C. Zach and M. Pollefeys, "Practical methods for convex multi-view reconstruction," *Lect Notes Comput Sc 6314*, 2010.

[136] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," *Symp Geom Process*, vol. 7, 2006.

[137] M. Waechter, N. Moehrle, and M. Goesele, "Let there be color! large-scale texturing of 3d reconstructions," *Proc ECCV*, 2014.

[138] Y. Wu, F. Tang, and H. Li, "Image-based camera localization: an overview," *Vis. Comput. Ind. Biomed. Art*, vol. 1, no. 8, 2018.

[139] P. Chen, C. Gong, A. Lewis, Y. Zhou, E. J. Seibel, and B. Hannaford, "Real-time flexible endoscope navigation within bladder phantom having sparse non-distinct features is enhanced with robotic control," in *SPIE Medical Imaging*, 2022.

[140] R. Reilink, A. M. Kappers, S. Stramigioli, and S. Misra, "Evaluation of robotically controlled advanced endoscopic instruments," *Int J Med Robotics Comput Assist Surg*, vol. 9, pp. 240–246, 2013.

[141] H. F. Talari, R. Monfaredi, E. Wilson, E. Blum, C. Bayne, C. Peters, A. Zhang, and K. Cleary, "Robotically assisted ureteroscopy for kidney exploration," R. J. Webster and B. Fei, Eds., vol. 10135. International Society for Optics and Photonics, mar 2017, p. 1013512. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2253862

[142] "Meshlab software," https://github.com/cnr-isti-vclab/meshlab/releases/tag/v2016.12, accessed: 2021-12.

[143] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, pp. 232–242, 2016.

[144] M. Niethammer, R. Kwitt, and F.-X. Vialard, "Metric learning for image registration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8463–8472.

[145] K.-N. Kim and R. Ramakrishna, "Vision-based eye-gaze tracking for human computer interface," in *IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 99CH37028)*, vol. 2. IEEE, 1999, pp. 324–329.

[146] B. Guenter, M. Finch, S. Drucker, D. Tan, and J. Snyder, "Foveated 3d graphics," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 6, pp. 1–10, 2012.

[147] E. Whitmire, L. Trutoiu, R. Cavin, D. Perek, B. Scally, J. Phillips, and S. Patel, "Eyecontact: scleral coil eye tracking for virtual reality," in *Proceedings of the 2016 ACM International Symposium on Wearable Computers*, 2016, pp. 184–191.

[148] L. Vaissie and J. Rolland, "Head mounted display with eyetracking capability," Aug. 13 2002, uS Patent 6,433,760.

[149] H. Hua, P. Krishnaswamy, and J. P. Rolland, "Video-based eyetracking methods and algorithms in head-mounted displays," *Optics Express*, vol. 14, no. 10, pp. 4328–4350, 2006.

[150] C. K. Sheehy, Q. Yang, D. W. Arathorn, P. Tiruveedhula, J. F. de Boer, and A. Roorda, "High-speed, image-based eye tracking with a scanning laser ophthalmoscope," *Biomedical optics express*, vol. 3, no. 10, pp. 2611–2622, 2012.

[151] E. D. Guestrin and M. Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections," *IEEE Transactions on biomedical engineering*, vol. 53, no. 6, pp. 1124–1133, 2006.

[152] A. Meyer, M. Böhme, T. Martinetz, and E. Barth, "A single-camera remote eye tracker," in *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*. Springer, 2006, pp. 208–211.

[153] A. Kar and P. Corcoran, "A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms," *IEEE Access*, vol. 5, pp. 16 495–16 519, 2017.

[154] T. A. Furness III and J. S. Kollin, "Virtual retinal display and method for tracking eye position," Nov. 13 2001, uS Patent 6,317,103.

[155] E. J. Seibel, R. S. Johnston, and C. D. Melville, "A full-color scanning fiber endoscope," in *Optical fibers and sensors for medical diagnostics and treatment applications VI*, vol. 6083. International Society for Optics and Photonics, 2006, p. 608303.

[156] T. D. Soper, M. P. Porter, and E. J. Seibel, "Surface mosaics of the bladder reconstructed from endoscopic video for automated surveillance," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 6, pp. 1670–1680, 2012.

[157] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8971–8980.

[158] P. Zarchan and H. Musoff, *Fundamentals of Kalman filtering: a practical approach*. American Institute of Aeronautics and Astronautics, Inc., 2013.

[159] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[160] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[161] P. Truong, S. Apostolopoulos, A. Mosinska, S. Stucky, C. Ciller, and S. D. Zanet, "Glampoints: Greedily learned accurate match points," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10 732–10 741.

[162] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[163] W. Li, L. Wang, W. Li, E. Agustsson, and L. V. Gool, "Webvision database: Visual learning and understanding from web data," *ArXiv*, vol. abs/1708.02862, 2017.

[164] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten, "Exploring the limits of weakly supervised pretraining," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 181–196.

[165] B. Frénay and M. Verleysen, "Classification in the presence of label noise: a survey," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 5, pp. 845–869, 2013.

[166] S. Hu, Y. Ying, X. Wang, and S. Lyu, "Learning by minimizing the sum of ranked range," *NeurIPS*, vol. 33, 2020.

[167] S. Hu, L. Ke, X. Wang, and S. Lyu, "Tkml-ap: Adversarial attacks to top-k multi-label learning," in *ICCV*, 2021, pp. 7649–7657.

[168] S. Hu, Y. Ying, X. Wang, and S. Lyu, "Sum of ranked range loss for supervised learning," *arXiv preprint arXiv:2106.03300*, 2021.

[169] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[170] T. Zhang and B. Yu, "Boosting with early stopping: Convergence and consistency," *The Annals of Statistics*, vol. 33, no. 4, pp. 1538–1579, 2005.

[171] D. Krueger, N. Ballas, S. Jastrzebski, D. Arpit, M. S. Kanwal, T. Maharaj, E. Bengio, A. Fischer, A. Courville, S. Lacoste-Julien *et al.*, "A closer look at memorization in deep networks," in *International Conference on Machine Learning (ICML)*, 2017.

[172] C. Lim, S. Han, and J. Lee, "Analyzing deep neural networks with noisy labels," in *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2020, pp. 571–574.

[173] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[174] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *International Conference on Machine Learning*, 2018, pp. 2304–2313.

[175] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in neural information processing systems*, 2018, pp. 8527–8537.

[176] H. Wei, L. Feng, X. Chen, and B. An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 726–13 735.

[177] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa, "Joint optimization framework for learning with noisy labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5552–5560.

[178] K. Yi and J. Wu, "Probabilistic end-to-end noise correction for learning with noisy labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7017–7025.

[179] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 4334–4343. [Online]. Available: http://proceedings.mlr.press/v80/ren18a.html

[180] Y. Shen and S. Sanghavi, "Learning with bad training data via iterative trimmed loss minimization," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5739–5748.

[181] I. Guyon, N. Matic, and V. Vapnik, "Discovering informative patterns and data cleaning," in *KDD Workshop*, 1994.

[182] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. Belongie, "Learning from noisy large-scale datasets with minimal supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 839–847.

[183] X. Liu, S. Li, M. Kan, S. Shan, and X. Chen, "Self-error-correcting convolutional neural network for learning with noisy labels," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 111–117.

[184] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[185] A. Menon, B. Van Rooyen, C. S. Ong, and B. Williamson, "Learning from corrupted binary labels via class-probability estimation," in *International Conference on Machine Learning*, 2015, pp. 125–134.

[186] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1944–1952.

[187] J. Goldberger and E. Ben-Reuven, "Training deep neural-networks using a noise adaptation layer," in *ICLR*, 2017.

[188] A. Ghosh, H. Kumar, and P. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 1919–1925.

[189] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Advances in neural information processing systems*, 2018, pp. 8778–8788.

[190] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," *arXiv preprint arXiv:1412.6596*, 2014.

[191] A. Kumar, A. Saha, and H. Daume, "Co-regularization based semi-supervised domain adaptation," *Advances in neural information processing systems*, vol. 23, pp. 478–486, 2010.

[192] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4320–4328.

[193] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2691–2699.

[194] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, and M. Sugiyama, "How does disagreement help generalization against label corruption?" in *36th International Conference on Machine Learning, ICML 2019*, 2019.

[195] J. Li, R. Socher, and S. C. Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," *arXiv preprint arXiv:2002.07394*, 2020.

[196] Y. Xu, P. Cao, Y. Kong, and Y. Wang, "L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise," in *Advances in neural information processing systems*, 2019, pp. 6225–6236.

[197] M. Elgendy, *Deep learning for vision systems*. Simon and Schuster, 2020, ch. 3.

[198] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[199] E. Ustinova and V. Lempitsky, "Learning deep embeddings with histogram loss," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[200] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen *et al.*, "Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Scientific data*, vol. 7, no. 1, pp. 1–14, 2020.

[201] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[202] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2024–2039, 2015.

[203] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2002–2011.

[204] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[205] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3828–3838.

[206] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[207] F. Mahmood and N. J. Durr, "Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy," *Medical image analysis*, vol. 48, pp. 230–243, 2018.